Genetic Structure of the Iraqi Population at 15 STRs and the Consequent Forensic Applications

By

Sarah D. Alden

Submitted to the graduate degree program in Anthropology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Master of Arts.

_____

Chair: Michael Crawford, PhD

_____

Jennifer Raff, PhD

_____

Majid Hannoum, PhD

Date Defended: 30 October 2017

The thesis committee for Sarah D. Alden certifies that this is the approved version of the following thesis:

Genetic Structure of the Iraqi Population at 15 STRs and the Consequent Forensic Applications

_____

Chair: Michael Crawford, PhD

Date Approved: 30 October 2017

Abstract

1061 individuals were sampled from the cities of Anbar, Baghdad, Basra, Diyala, Najaf, and Wasit in Iraq and typed for 15 forensic STRs to explore the genetic structure of Iraq and develop a forensic DNA database. Analyses found that Iraq is similar to other countries in the Middle East, particularly Iran and Turkey, and is more similar to Europe than either Asia or Africa. Iraq is genetically diverse; a clustering algorithm was used to infer the number of genetic clusters in the population and the best fit was eight genetic clusters. Baghdad provides a good representation of the rest of country while Anbar is the most genetically distinct. This may be because Anbar is the only city sampled in a Sunni-dominant region. Although genetic structure differs significantly between the cities, most of the genetic differentiation is between genetic clusters rather than cities.

These loci had an average heterozygosity of 0.779, homozygosity of 0.221, polymorphism information content of 0.77, power of discrimination of 0.927, and power of exclusion of 0.563. At these loci, a matching genotype will occur, on average, in 1 in 8.152 x 10^{17} individuals. For paternity tests, the average paternity probability for a matching profile is 99.9997%. For both measures, this can be taken as an exact match. These loci are appropriate for use in forensic and paternity testing for this population.

Acknowledgments

I would like to thank the people of Iraq who generously donated their DNA for the project. I must also thank the researchers at the Forensic DNA Center for Research and Training at Al-Nahrain University who collected the samples and performed all the laboratory work and typing: Majeed Arsheed Sabbah, Mohammed Mahdi Al-Zubaidi, Haider K. Alrubai, Dhuha Salim Namaa, Thoalnoon Younis Saliha, Hala Khalid Ibrahem, Salba Jaber Al-Awadi, and Adnan Issa Al-Badran, as well as Basim Muften at the Iraqi Ministry of Interior.

I would like to thank my advisor and committee chair, Dr. Michael Crawford, for his support and guidance as well as my committee members, Dr. Jennifer Raff and Dr. Hannoum. I would like to acknowledge the help of Dr. Kristine Beaty, Amanda Kittoe, and Michael Guarino in interpreting various analyses and providing feedback on my thesis, and Corinne Butler for making sure that I knew what needed to be done and when, as well as the Department of Anthropology and the Laboratories of Biological Anthropology. Finally, I would like to thank my family for their support.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Iraq is a population that needs continued study in anthropological genetics. There have been relatively few English-language publications on the genetics of Iraq. The genetic structure has not been well established for certain regions of the country, particularly western Iraq, nor have different regions really been compared to each other to see if genetic structure differs among the different regions. This study seeks to look at whether genetic structure varies between six different cities in Iraq and then to create a forensic DNA database for the country of Iraq.

Population genetics looks to study the genetic variation within populations. To do this, researchers measure the frequencies of alleles in a population and look at changes in these frequencies over time and/or location. Any pattern found in the genetic makeup of a population is the genetic structure of that population. These patterns allow for inference about the genetic makeup of a single individual by studying other members of the same population. Forensic genetics involves applying genetic structure to identify individuals for legal purposes.

Iraq is a country in the Middle East surrounded by Syria to the west, Jordan to the southwest, Saudi Arabia to the south, Kuwait to the southeast, Iran to the east, and Turkey to the north. Baghdad is the capital and largest city. The main ethnic groups are Arabs (75-80%) and Kurds (15-20%), with some additional minority populations (Kirmanj, 2013; The World Factbook, 2017). Kurds are concentrated in northern Iraq while the rest of the country are mainly Arab. However, ethnicity is a difficult term to define and definitions vary according to who is using them. It can refer to a person's ancestry but it can also refer to, or include, cultural groups (Eriksen, 2010). It is a term used to refer to relationships, or aspects of relationships, between groups that are considered, by themselves and others, as distinctive. In social anthropology, this distinction is cultural; however, ethnic groups tend to have, or are believed to have, a common

descent which relates ethnicity to biology (Eriksen, 2010). With regards to Iraq, the concept of ethnicity is tied to language. Arabs are people who speak Arabic, as opposed to Kurdish or Farsi, but whether that relates to genetic differences between the groups remains to be seen; however, there is evidence that these are genetically related groups (Al-Zahery, et al., 2003; Nebel, et al., 2001). 97-99% of the population are Muslim with ~40% Sunni and ~60% Shia (Kirmanj, 2013; The World Factbook, 2017). Among the Muslim Arab population, ~25% are Sunni while ~75% are Shia while the Kurdish population is mainly Sunni (Kirmanj, 2013). Among Sunni Muslims are mainly found in western and northern Iraq while Shia Muslims are mainly found in eastern and southern Iraq (Kirmanj, 2013; The World Factbook, 2017).

First-cousin marriage is common in Iraq with one study in northern Iraq reporting a consanguinity rate of 27.2% (second-cousin or closer) leading certain regions to require pre-marital testing for hemoglobinopathies (Al-Allawi, et al., 2015) as well as several reports in the literature of recessive diseases linked to consanguinity (Donaldson, Tucket, & Grant, 1980; Hamamy & Al-Hakkak, 1989; Hamamy, Makrythanasis, Al-Allawi, Muhsin, & Antonarakis, 2014; Henningsen, Svendsen, Lildballe, & Jensen, 2014). Inbreeding can also affect genetic structure by 1) increasing genetic differences between groups in the absence of gene flow, and 2) decreasing the genetic variation within the group by increasing homozygosity.

Y-chromosome analyses revealed little difference between Iraqi Arabs, Assyrians, and Kurds (Al-Zahery, et al., 2003) and, while there is minimal gene flow between the Jewish and Muslim populations in Iraq, both populations are more genetically similar to each than either are to populations in Europe (Nebel, et al., 2001). Mitochondrial studies found Iraq to be most similar to Iran (Al-Zahery, et al., 2003) and Turkey (Farzad, et al., 2013; Tomas, Diez, Moncada, Borsting, & Morling, 2013) and generally similar to other countries in the Levant, but

significantly different from countries in the Arabian Peninsula (Al-Zahery, et al., 2003). Both mtDNA and Y-chromosome revealed expansions out of Iraq into Europe as well as gene flow from Europe, Central Asia, and Africa with the most gene flow from Europe and the least from Africa (Al-Zahery, et al., 2003) and the gene flow from Africa being mainly female (Richards, et al., 2003). There have been some studies focused on specific populations within Iraq. These studies will be reviewed in greater depth in Chapter 2: Literature Review.

However, the studies currently in the literature do not really explore the genetic structure of the general population in Iraq among different regions. Most of the studies are focused in a single region in Iraq, usually Baghdad, or from Iraqi immigrants living in other countries. This study sampled from six different cities in Iraq located in different regions and provides an opportunity to look at whether genetic structure differs between the cities/regions. And, if there is genetic differentiation among the cities, can any inferences be made as to the forces that are causing it?

This study looked at the genetic structure of Iraq and form a forensic DNA database using 15 autosomal STRs which are commonly used for forensic purposes: D8S1179, D21S11, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, D2S1338, vWA, TPOX, D18S51, D5S818, and FGA. It compared the allele frequencies of these loci to surrounding countries (Syria, Iran, Turkey, Kuwait, and Saudi Arabia) and to countries in Europe (Poland and Belgium), Africa (Equatorial Guinea and Angola), and Asia (China and Japan).

1061 samples were collected from the cities of Anbar, Diyala, Basra, Wasit, Najaf, and Baghdad. Baghdad and Najaf are in mid-eastern Iraq, Diyala and Wasit in the east, Basra in the southeast, and Anbar more in western Iraq (Figure 1). Genetic structure was explored using common statistical analyses: Hardy-Weinberg equilibrium, F-statistics, Nei's genetic distance,

analysis of molecular variance, principal component analysis, multidimensional scaling, and discriminant analysis of principal components. To create the forensic DNA database, each locus had its allele frequency, heterozygosity, homozygosity, matching probability, power of discrimination, polymorphism information content, power of exclusion, and typical paternity index calculated for each city. These analyses will be explained more fully in Chapter 3: Methods.

Some limitations of this study are that the dataset is not a random sample of the Iraqi population, and the data are de-identified so there is no information on the individuals' ethnicities, religions, of places of birth. Inferences will be made using known demographics of the regions under study. The data were treated as though the city in which a sample was collected was the individual's place of birth. The assumption was also made that most the samples are from Arab individuals with the possibility of some minority populations which are most likely to be found in Baghdad. Another limitation is that no samples were taken from northern Iraq or the extreme south of the country, and only one city, Anbar, represents western Iraq. However, if there is genetic differentiation among the regions included in the study then a next step would be to further explore by including more dispersed regions.

Figure 1. A map of Iraq with cities marked (Google Maps, 2017).

Chapter 2: Literature Review

A History of Iraq

Modern-day Iraq sits right inside ancient Mesopotamia, part of the Fertile Crescent which was one of the cradles of civilization (Stansfield, 2007). The floodplains of the Tigris and Euphrates Rivers in what is now southern and central Iraq made cultivation of crops possible. In certain areas, Emmer wheat and barley grew naturally and could be easily harvested without any need for planting, and proto-Neolithic people took advantage of this resource (Foster & Foster, 2009; Roux, 1992). However, this fertile area was only in narrow bands adjacent to the rivers while beyond that was arid land with localized rainfall and extreme temperature changes which likely necessitated the development of irrigation (Foster & Foster, 2009; Stansfield, 2007).

The earliest traces of a human presence in what is now known as Iraq are "pebble tools" that were found north of present-day Mosul in the upper Tigris River valley (Roux, 1992). These tools have been classified as part of the Upper Acheulean industry which would place them in the Lower Paleolithic. Other important Paleolithic sites are in northern Iraq and the Zagros Mountains. Mesolithic cemeteries were found in Shanidar Cave in the Zagros Mountains (Roux, 1992).

During the Neolithic, humans in Iraq began practicing agriculture, domestication of animals, and settled into mud houses (Roux, 1992). Neolithic sites have revealed square dwellings made from pressed mud and containing multiple rooms, mud ovens, and baked-in clay basins that were sunk into the ground. Sites have yielded bone spoons and needles as wells as stone spindle whorls. Evidence of agriculture and domestication of animals have been found in the form of carbonized grains of barley and wheat, and the bones of domesticated sheep, cattle, pigs, and dogs (Roux, 1992).

The post-Neolithic, pre-historical era in Iraq can be split into various periods, which are characterized by their different styles of pottery (Roux, 1992). The first of these is the Hassuna period (*c*. 5800-5500 B.C.) which featured small villages made up of mud dwellings surrounding open courts. This was followed by the Samarra period (*c*. 5600-5000 B.C.), which provides the first signs of irrigation in Iraq. This period overlaps somewhat with the Halaf period (*c*. 5500-4500 B.C.). These people were farmers and pastoralists; they grew a variety of grains, lentils, and other vegetables as well as keeping sheep, goats, pigs, and cattle. It has been hypothesized that ranked social classes existed during this period. The Halafian culture was gradually replaced by the Ubaid culture (*c*. 5000-3750 B.C.) in which we see more intensive irrigation and a movement towards urbanization. The presence of obsidian, gold, and amazonite reveals that long-distance trade occurred during this time (Roux, 1992).

The Ubaid culture appears to have given rise to the Uruk culture (*c*. 3750-3150 B.C.) since there is no clear break between them or evidence of invasion (Roux, 1992; Stansfield, 2007). This period was followed by the Jemdat Nasr period (*c*. 3150-2900 B.C.). The Uruk period saw the emergence of urbanization in Iraq, intensive irrigation and use of ploughs, the gradual appearance of cuneiform writing, and the first established legal systems (Roux, 1992; Stansfield, 2007). There is evidence that the Jemdat Nasr period settlements were very organized with centralized building, administrative cuneiform tablets, and cylinder seals (Roux, 1992).

Sumerian city-states arose during the third millennium B.C. (Stansfield, 2007). These were large urban centers that were ruled by hereditary dynasties. They were not a unified political entity, but rather each city-state was independent. It was during this period that cuneiform writing flourished and there were developments in mathematics and astronomy. The city-states fell to Sargon around 2370 B.C. who unified Mesopotamia under his rule. He

established the Akkadian dynasty which continued for approximately 200 years until it collapsed following chronic rebellions. The Ur dynasty began around 2100 B.C. and established a vast trading network (Stansfield, 2007).

The Ur dynasty fell and what followed is known as the "Old Babylonian" period (*c*. 2000-1600 B.C.) (Stansfield, 2007). During this time, there were two main power centers: Babylon in the north and Larsa in the south. The two regions were unified under the Babylonian king, Hammurabi. This period introduced taxation and military conscription. Babylon was sacked by the Hittites in 1595 B.C. and there was a period of instability until the Assyrians, a group of Semitic people from northern Mesopotamia, took power (Stansfield, 2007).

The Assyrian state ruled from *c*. 1300-600 B.C., and controlled the region from Iran to Egypt (Stansfield, 2007). The state eventually fell in 609 B.C. after a series of attacks from an alliance formed of the Babylonians from the south and Medes from the north. This resulted in the "New Babylonian" period with the Babylonians taking over Mesopotamia and the Medes taking over the Zagros Mountains and the Iranian plateau. The "New Babylonian" period continued until Cyrus the Great, originally a vassal of the Median king, established a power base in Persia, formed an alliance with Persian tribes, defeated the Medes, and conquered Babylon. By 486 B.C., the Persian empire ranged from Macedon to Egypt and from Mesopotamia to India (Stansfield, 2007).

By this time, the population in Iraq was quite diverse (Stansfield, 2007). Semi-autonomous Arab tribes roamed the deserts, the Kurds (believed to be descendants of the Medes) lived in the northern mountains, and there were pockets of Greeks, Indians, and Africans scattered throughout the region. The main religion was Nestorianism, but there were also Jewish

communities throughout the region and Zoroastrianism was popular among the Persian and Kurdish people (Stansfield, 2007).

Mohammed established the basis for the religious teachings of Islam in the Arabian town of Mecca (in modern Saudi Arabia) at the end of the 6th century A.D. (Stansfield, 2007). Due to opposition to his teachings in Mecca, he and his followers fled to Medina (the *hijrah*) in 622 A.D. and founded the Islamic state. The state used the native tribal structure of the Arabs to bring social and political organization to Mohammed's teachings and Islam spread mainly through voluntary conversion of Arab tribes with occasional battles with Persian forces. By the time Mohammed died in 632 A.D., the Islamic state dominated the Arabian peninsula and was moving into Mesopotamia. A decisive battle was fought in 637 A.D. at Qadisyyah with the Muslims defeating the Persians. The Persians retreated, and Mesopotamia and the northern Zagros Mountains were absorbed into the Islamic state (Stansfield, 2007).

Following Mohammed's death, there was a power struggle between two Islamic groups which would ultimately split the religion in two (Stansfield, 2007). The first group, which became known as Sunnis, wanted to elect a caliph by consensus among the community leaders while the second group, now known as the Shia, believed that the caliph should be related to Mohammed and pushed for his son-in-law, Ali, to be caliph. Eventually, decades later, Ali was finally recognized as the fourth caliph of the Rashidun Caliphate; however, he was assassinated in 661 A.D., a mere five years later (Stansfield, 2007; Hourani, 1991).

Muawiya, a close relative of the third caliph, Uthman ibn 'Attan, made a claim for the caliphate and Ali's eldest son, Hassan, gave up the caliphate to Muawiya (Hourani, 1991; Spuler, 1994). Muawiya established the Umayyad Caliphate and moved the capital to Damascus, Syria (Hourani, 1991). When Muawiya died in 680 A.D., the supporters of Hussein, Ali's second son,

attempted to secure the caliphate for him. But Muawiya's son, Yazid, had his forces kill Hussein at Karbala, in modern-day Iraq, and Yazid succeeded his father as caliph. However, the Shia community continued to regard Hussein's descendants as their leaders, known as imams (Stansfield, 2007).

The Umayyad ruling family was replaced by Abbasid family in 750 A.D. (Hourani, 1991; Stansfield, 2007). The Abbasids had the advantage of being descended from one of Mohammed's uncles, 'Abbas. Also, the family had held a prominent position in pre-Islamic Mesopotamia and had a strong following both in Persia and Khurasan. They overtook the Umayyad family in 750 A.D. and established the Abbasid Caliphate (Hourani, 1991).

During the Abbasid rule, the population of Iraq grew to 20 million and the city of Baghdad was established, but in 945 A.D. Iraq once again fell under Persian rule though the Caliphate remained as ceremonial religious figures (Stansfield, 2007). In 1258 A.D., the city of Baghdad fell to Mongol invaders. Iraq was then under the control of the Mongol Il-Khanate until it fell to Timur of Samarkand (in modern-day Uzbekistan) in the late 14th century. Iraq was then seized by the Qara Qoyunlu, "Black Sheep", a federation of Turkmen tribes which were succeeded by a rival Turkmen federation, the Aq Qoyunlu, "White Sheep" (Stansfield, 2007).

At this time, Iraq was considered a geopolitical prize for two rival powers: the Shia Safavids of Persia and the Sunni Ottomans of Anatolia (Stansfield, 2007). The Safavids took control of the Aq Qoyunlu capital of Tabriz (in modern-day Iran) in 1501 and Baghdad in 1508. However, their rule was short-lived as the Ottomans took control of Azerbaijan in 1514, Northern Iraq shortly thereafter and, finally, Baghdad in 1534. Iraq then remained under Ottoman rule for nearly four centuries (Stansfield, 2007).

The Ottoman Empire was dissolved following World War I, when it acted as one of the Central Powers (including Germany, Austria-Hungary, and Bulgaria) and was defeated by the Allied Powers (Stansfield, 2007; Tripp, 2002). The Iraqi province of Basra was occupied by British forces in 1914 (Stansfield, 2007; Tripp, 2002). By 1917, the city of Baghdad had fallen to Britain which was followed by Mosul, in northern Iraq, in 1918 (Tripp, 2002).

The San Remo Conference was held in April of 1920 to partition out the territories of the Ottoman Empire, with Britain receiving the mandates for Palestine and Iraq (Stansfield, 2007; Tripp, 2002). In March of 1921, the Cairo Conference was held by British officials to discuss the Middle East and make policy decisions. They decided to establish a monarchy to run Iraq while they kept the mandate. They appointed Emir Faisal as the king of this monarchy. In August of 1921, he was installed as King Faisal I and in 1924 the Anglo-Iraqi Treaty was ratified to conclude the resolution made at the Cairo Conference to establish the Hashemite Kingdom in Iraq (Stansfield, 2007; Tripp, 2002).

In 1925, the Iraqi government signed the Turkish Petroleum Company (TPC) oil concession which allowed the TPC to explore for oil in Iraq in exchange for the Iraqi government receiving a royalty for any oil that was extracted (Tripp, 2002). Because of the discovery of oil in the region following World War I, both Turkey and Great Britain wanted the Mosul Vilayet. Vilayets were the major administrative districts within the Ottoman Empire. The Treaty of Lausanne in 1923 failed to reach an agreement on this territory, so the matter was sent to the League of Nations for a final decision. Ultimately, they made the Mosul Vilayet part of Iraq, and in 1927 the British struck oil in Kirkuk, a city within the Mosul Vilayet (Stansfield, 2007; Tripp, 2002).

Following the Anglo-Iraqi treaty of 1930, the Kingdom of Iraq was granted formal independence in 1932 (Stansfield, 2007; Tripp, 2002). King Faisal I died on September 8, 1933 and was succeeded by his son, Ghazi. During King Ghazi's rule, Iraq experienced its first military coup. General Bakr Sidqi overthrew Prime Minister Yasin al-Hashimi in 1936, installing Hikmat Sulaiman as the new Prime Minister, and Sidqi himself as Chief of Staff. Bakr Sidqi was assassinated the following year in Mosul. King Ghazi died in a car crash in April of 1939. Ghazi's four-year-old son was crowned King Faisal II with his uncle, Abd al-Ilah, appointed as regent (Stansfield, 2007; Tripp, 2002).

World War II commenced in 1939 with Iraq formally aligned with the Allied Powers (Stansfield, 2007). However, the Golden Square, a group of Sunni Arab nationalist military officers, wanting to be rid of British power in Iraq, sought support from Germany and staged a coup d'état in April 1941 (Stansfield, 2007; Tripp, 2002). Regent Abd al-Ilah and Prime Minister Nuri al-Sa'id fled Iraq, and Rashid Ali al-Gaylani was installed as Prime Minister. Great Britain responded by sending troops into Basra (southeastern Iraq) and moving north into Baghdad. By May, al-Gaylani and the members of his government fled the country while the mayor of Baghdad negotiated an armistice with Great Britain. Regent Abd al-Ilah, Prime Minister Nuri al-Sa'id, and others who had fled during the coup returned to Iraq (Stansfield, 2007; Tripp, 2002).

In May of 1953, the regency ended as King Faisal II came of age and was enthroned (Tripp, 2002). In February of 1955, the Baghdad Pact was formed by Iraq, Iran, Pakistan, and Turkey, and endorsed by Great Britain. Modeled after the North Atlantic Treaty Organization (NATO), the participating members of the Baghdad Pact agreed to cooperation, protection, and non-involvement in each other's affairs as well as to prevent Soviet expansion into the Middle East by providing a strong front at the southwestern USSR border (Tripp, 2002).

Egypt denounced Iraq for having formed a pact with Turkey and Iraqi Prime Minister Nuri al-Sa'id was criticized in Iraq by the pan-Arab opposition for continuing Iraq's "enslavement" to Western power (Tripp, 2002). When British, French, and Israeli forces bombed Egypt during the Suez Crisis in 1956, Iraq, as a British ally, supported the invasion which was met with widespread disapproval among the Iraqi populace who largely sided with Egypt. Protests and demonstrations were held in cities throughout Iraq. The Suez Crisis bolstered the pan-Arab cause and undermined the pro-Western government (Tripp, 2002).

Civil unrest ultimately led to the 1958 coup d'état (Stansfield, 2007; Tripp, 2002). On July 14, 1958, a secret military group, known as the Free Officers, led by Brigadier Abd al-Karim Qasim and Colonel Abd al-Salam Arif, ordered troops into the capital where they took over the Royal Palace and executed King Faisal II and Prince Abd al-Ilah. Prime Minister Nuri al-Sa'id's house was also raided. Although he managed to escape, he was captured and shot the following day. The Iraqi Republic was established with Abd al-Karim Qasim as Prime Minister (Stansfield, 2007; Tripp, 2002).

Kurdish leader, Mulla Mustafa Barzani, who had been exiled to the USSR was allowed to return to Iraq in 1959 which lead to a revival of the Kurdistan Democratic Party (KDP) (Stansfield, 2007; Tripp, 2002). Arif, who advocated joining the United Arab Republic (UAR), was ordered by Qasim to act as ambassador to Germany. Arif left but tried to return to Baghdad in secret and was arrested (Tripp, 2002).

Qasim's refusal to join to UAR and his alliance with the ICP led the Ba'ath party, including 22-year-old Saddam Hussein, to make an unsuccessful assassination attempt on October 7, 1959 (Tripp, 2002). Qasim's government became increasing dictatorial. His

relationship with the ICP broke down and he refused to grant them a license to organize as a legal political party. In December 1959, Iraq withdrew from the Baghdad Pact (Tripp, 2002).

In 1961, the Kurdish Democratic Party (KDP), lead by Mulla Mustafa Barzani, proposed a plan to establish Kurdish autonomy which was rejected by Qasim (Stansfield, 2007; Tripp, 2002). This led to Kurdish rebellions against Qasim and Kurdish support for the Ba'ath party military coup that occurred on February 8, 1963. Qasim and his supporters were captured, brought before a tribunal of Ba'athist and pan-Arab officers, and executed. Abd al-Salam Arif, who was not a Ba'athist but who had a considerable following in the military, was instituted as president while his vice president and prime minister was the Ba'athist Ahmad Hasan al-Bakr. Almost immediately, the regime broke into factions with differing ideas about the identity of Iraq and the direction it should go, and on November 18, 1963 President Arif removed members of the Ba'ath party from positions of power (Stansfield, 2007; Tripp, 2002).

On April 18, 1966, President Arif died in a helicopter crash and his brother, Abd al-Rahman Arif, succeeded him as president (Stansfield, 2007; Tripp, 2002). A second Ba'athist coup d'état, led by former prime minister Ahmed Hassan al-Bakr, occurred on July 17, 1968. al-Bakr was established as president and Arif was sent into exile. President al-Bakr quickly removed important military figures, communists, and Kurdish nationalists from government to remove any possible internal threats to the new regime (Stansfield, 2007). In November of 1969, Saddam Hussein was appointed the deputy chair of the Revolutionary Command Council (RCC) which was the preeminent decision-making body in the state (Stansfield, 2007; Tripp, 2002).

Mustafa Barzani had attempted peace negotiations with both President Abd al-Salam Arif and his brother, but no agreement was met and by April of 1965 there was full-scale war in Kurdistan (Tripp, 2002). Eventually, the Iraqi-Kurdish Autonomy Agreement of 1970 was

reached on March 11, 1970 to end the Kurdish-Iraqi War and create an autonomous region. However, fighting in Kurdistan resumed by 1972 and by 1974 the negotiations had collapsed, the Kurds rebelled with Iranian support, leading to the Second Kurdish-Iraqi War (Tripp, 2002).

By 1974, Saddam Hussein was vice president of the RCC and second in power only to President al-Bakr (Stansfield, 2007). In 1975, Iraq entered into the Algiers Agreement with Iran to end their border disputes and remove Iranian support from the Kurds (Tripp, 2002). Without Iran, the Kurdish revolt collapsed, and disagreements within the Kurdish movement lead to a split in leadership with Masoud Barzani, Mustafa Barzani's son, leading the KDP-Provisional Leadership and Jalal Talabani leading the Popular Union of Kurdistan (PUK) (Tripp, 2002).

On July 16, 1979, President al-Bakr resigned and Saddam Hussein succeeded him (Stansfield, 2007; Tripp, 2002). Territorial arguments between Iraq and Iran and fears that the 1979 Iranian Revolution would inspire revolution within Iraq led to the Iran-Iraq War which lasted from 1980 to 1988. Although Iraq declared itself victorious, the economy was devastated and no territory was gained as Iran and Iraq retained their pre-war borders. Two years later, border disputes with Kuwait as well as disagreements over Iraq's financial debt to Kuwait led to Iraq invading Kuwait on August 2, 1990. The United Nations adopted Security Council Resolution (SCR) 660 which demanded that Iraq withdraw from Kuwait immediately. In response to Iraq's refusal to comply with SCR 660, the UN adopted SCR 661 on August 6 which placed international sanctions on Iraq, prohibiting import and export of products and commodities except for medical and/or humanitarian purposes. In January of 1991, the US Senate authorized use of force and on January 17 Operation Desert Storm began. It lasted until a ceasefire on February 28. On April 3, 1991, the UN adopted SCR 687 which provided the terms for the ceasefire and required Iraq to terminate all weapons of mass destruction programs. In

1992, northern Iraq became a *de facto* autonomous region known as Iraqi Kurdistan (Stansfield, 2007; Tripp, 2002).

Throughout the 1990s, Iraq suffered from increased poverty and undernutrition resulting in the adoption of UN SCR 986 which would provide humanitarian aid to the country and allow Iraq to export oil (Tripp, 2002). Although it was adopted by the UN on April 14, 1995, Iraq did not accept SCR 986 until February of 1996. The first oil exported as part of SCR 986 was in December of 1996 marking Iraq's return to the world oil market (Tripp, 2002).

On September 11, 2001, Al-Qaeda attacked the US and there was speculation about Iraq's involvement in the attacks (Stansfield, 2007). In September of 2002, Iraq allowed inspectors from the United Nations Monitoring, Verification and Inspection Commission (UNMOVIC) into the country. They issued a report stating that Iraq's capability for producing WMDs had not changed since 1992 when the bulk of their arsenals were dismantled as part of the terms of the Desert Storm ceasefire and that they may still be hiding WMDs. On November 8, 2002, the UN adopted SCR 1441 which stated that Iraq was in breach of the ceasefire agreement of SCR 687 and demanded full compliance with UNMOVIC. In November of 2002, UNMOVIC inspectors re-entered the country and in March of 2003 they issued a report that neither confirmed nor denied Iraq being in violation of SCR 1441 and recommended further inspections. However, on March 20, 2003, the US began an air attack on Baghdad while US ground troops invaded southern Iraq. The US took control over Baghdad on April 9. Saddam Hussein went into hiding, but was captured on December 13, 2003 (Stansfield, 2007).

In January of 2005, elections were held for the Transitional National Assembly (TNA) (Stansfield, 2007). In April of 2005, the TNA selected Jalal Talabani as president of Iraq and Ibrahim al-Ja-afari as prime minister. In October of 2005, Saddam Hussein's trial began. A new

constitution was drafted and ratified, and on December 15, 2005 Iraqi voted for the first full-term parliamentary government since the US invasion. The Shia-led United Iraqi Alliance was declared the winner of the election; however, a significant portion of the Sunni community did not vote after Al-Qaeda issued warnings not to vote in the elections and that those who did would be legitimate targets. On June 8, 2006, it was announced that Abue Musab al-Zarqawi, the Al-Qaeda leader in Iraq, had been killed in an air strike (Stansfield, 2007). Saddam Hussein was convicted of crimes against humanity on November 5, 2006 and executed on December 30, 2006 (BBC News, 2006, December 30).

The inability of Iraqi soldiers to prevent looting and violence following the withdraw of British troops led to doubt over whether the government would be able to maintain control if US troops handed over security to them (BBC News, 2006, July 13; Washington Post, 2006, August 25; Washington Post, 2006, August 26). Attacks from insurgents resulted in tens of thousands of deaths throughout Iraq (BBC News, 2006, August 28; CNN, 2006, November 12; NY Times, 2006, October 20). In January of 2007, President Bush deployed an addition 30,000 troops to Iraq (Time, 2008, January 31) while in February 2007, UK Prime Minister Tony Blair announced that they would be withdrawing troops from Iraq (BBC News, 2007, February 21).

In May of 2007, Iraq endorsed a bill that stipulated that there be no increase in the number of foreign troops within Iraq and demanded a timetable for troop withdrawal (Associated Press, 2007, May 10). US troops started to withdraw from Iraq in June of 2009 (BBC News, 2009, June 30) and the last of the US troops exited Iraq through Kuwait in December of 2011 (CNN, 2011, December 18). On December 18, 2012, Iraqi president Jalal Talabani suffered a stroke (BBC News, 2012, December 18) and in July of 2014 Fuad Masum was elected (World Bulletin, 2014, July 24). Fuad Masum nominated Haider al-Abadi to succeed Nuri al-Maliki as

prime minister (The Guardian, 2014, August 11; BBC News, 2014, September 9; CNN, 2014, August 11).

Insurgent violence in Iraq increased with protests against the government and militant groups attacking and taking over Iraqi cities, including a significant portion of Anbar (BBC News, 2012, December 28; BBC News, 2014, June 18; BBC News, 2013, January 2) and northern Iraq (Stansfield, 2016). However, ISIS territory continues to shrink as the Iraqi government has reclaimed many cities and towns including the major city of Mosul (NY Times, 2017, September 2). Fighting between ISIS and the Iraqi governments continues. There is also currently discord between the Iraqi government and the Kurdistan regional government following Iraqi Kurdistan's vote for independence on September 25, 2017 (NY Times, 2017, September 26).

Population Genetics

Inbreeding

Consanguinity is high in Iraq with first-cousin marriage a common occurrence. A study examining the couples that visited the Duhok premarital-screening center from 2008 to 2012 found the consanguinity (second-cousin or closer) rate among couples was 27.2% (Al-Allawi, et al., 2015). Among the Karaite community that immigrated from Iraq, the inbreeding coefficient ranged from 0.012 to as high as 0.053 (Goldschmidt, Fried, Steinberg, & Cohen, 1976). This high rate of consanguinity has resulted in various recessive diseases presenting within families. A 1989 study found a high correlation between consanguinity and congenital malformations with severely disturbed reproductive health among families with an inbreeding coefficient above 0.03 (Hamamy & Al-Hakkak, 1989). High consanguinity rates have also been found in Iraqi families suffering from hypohidotic ectodermal dysplasia (Henningsen, Svendsen, Lildballe, & Jensen,

2014), recessive thrombocytopenia (Hamamy, Makrythanasis, Al-Allawi, Muhsin, & Antonarakis, 2014), and growth hormone deficiency (Donaldson, Tucket, & Grant, 1980).

This high rate of consanguinity has also resulted in hemoglobinopathies becoming a major health concern within Iraq to the point that certain provinces in northern Iraq have passed legislation requiring that couples undergo pre-marital screenings for genetic mutations (Al-Allawi, et al., 2015). β-thalassemia is found at uniform rates across Iraq with an average carrier rate around 4%, while sickle-cell anemia rates vary from 0% in the middle of the country to as high as 16% in the extreme north and south. In addition to pre-marital screening, chorionic villus sampling is offered to pregnant couples who are known carriers to test the fetus for hemoglobinopathies (Al-Allawi, et al., 2015).

Y Chromosome

According to a study in 2003, the most represented haplogroup in Iraq was J-12f2 at 58.3%. This is one of the highest reported frequencies of this haplogroup. It has also found at high incidence in the Zagros Mountains of Iran and in Syria. It is also found at decreasing frequency from SE Europe to NW Europe (Al-Zahery, et al., 2003). It has been proposed that the haplogroup arose in the Fertile Crescent and spread into western Europe (Al-Zahery, et al, 2003). There is also support for expansion of J1e from the Fertile Crescent into the arid Arabian Peninsula during the Neolithic as evidenced by a southward decreasing trend and coalescence dates of around 10 kya (Abu-Amero, et al., 2009; Chiaroni, et al., 2010). The next most frequent haplogroups were Eurasian, R-M269 (10.8%) and R-M17 (6.5%) which indicate gene flow from Central Asia/Eastern Europe (Al-Zahery, et al., 2003). There is also gene flow from African as evidenced by the presence of E-M35 (10.8%), and haplogroup I (0.7%) shows some limited gene flow from Europe (Al-Zahery, et al., 2003)..

Overall, the Y-chromosome analysis supports the hypothesis that haplogroup J1 originated in the Eastern Fertile Crescent before expanding into Europe and the Arabian peninsula (Abu-Amero, et al., 2009; Al-Zahery, et al., 2003; Chiaroni, et al., 2010) followed by subsequent gene flow with Europe, Central Asia, and Africa which is not surprising given the region's position as a cross-road between the continents and its history which includes lots of movements of people through the region (Al-Zahery, et al., 2003).

Within the Marsh Arabs, a group of people who traditionally lived in the marshlands between the Tigris and Euphrates Rivers, only a small portion of their Y-chromosome gene pool is the result of gene flow (Al-Zahery, et al., 2011). Unlike the general Iraqi population, the Marsh Arabs had essentially no contribution from western Eurasia or Africa, very little from Northern Middle East, with the most from Southwest Asia. Haplogroup J accounted for 84.6% of the Marsh Arab Y-chromosome gene pool with almost all belonging to the J1-M267 clade. Both haplogroups E and R1 were found at a lower frequency in Marsh Arabs than in the general Iraqi population (6.3% and 2.8%, respectively). There were low frequencies of Q (2.8%), G (1.4%), L (0.7%), and R2 (1.4%). Y-chromosome heterogeneity was much lower in Marsh Arabs (H = 0.461) than in the general populations (H = 0.887). The authors believe that the Y-chromosome evidence supports the Marsh Arabs as being autochthonous to the region (Al-Zahery, et al., 2011).

An exploration in the Y-chromosome relationships among the Iraqi Jewish populations found that the Y-chromosomes of Kurdish Jews were genetically closer to other Jewish populations (Sephardim and Ashkenazim) than to the Muslim Kurds (Nebel, et al., 2001). Additionally, the Y-chromosomes in the Sephardic community in Iraq were found to be very closely related to the Sephardic community in North Africa and to the Kurdish Jews. According

to these findings, any male gene flow that has occurred between the Jewish populations in Iraq and their Muslim neighbors has been below the level of detection (Nebel, et al., 2001).

However, the Y-chromosomes of Iraqi Jews, Kurds, and Arabs were all found to be genetically closer to each other than any of the groups are to Europeans, so even though the Jewish and Kurdish population have maintained themselves as isolates, they are still a fundamental element of the genetic make-up of the region (Nebel, et al., 2001). This agrees with findings using classical markers which found that Kurds tend to cluster very closely with Iraqi Arabs (Cavalli-Sforza, Menozzi, & Piazza, 1994), but differs from what has been found using mtDNA which found Kurds to be more closely related to Europeans than to Middle Easterners (Comas, Calafell, Bendukidze, Fañanás, & Bertranpetit, 2000). Simiarly, Y-chromosome analysis did not show any significant difference between Iraqi Arabs, Assyrians, and Kurds (Al-Zahery, et al., 2003).

It was also found that there is high genetic affinity across major language divisions (Indo-European, Semitic, and Turkic) within the region which was interpreted as indicating that the Middle Eastern Y-chromosome pool predates the emergence and/or introduction of the different languages found in the region (Nebel, et al., 2001). Additionally, a study that compared frequencies of 17 Y-STRs in Turkey to surrounding countries (including Iraq) using AMOVA found no significant difference between Iraq and Turkey (Ozbas-Gerceker, Bozman, Arslan, & Serin, 2013).

Mitochondrial DNA

In a study of 216 unrelated males from Baghdad (178 Arabs, 25 Assyrians, and 13 Kurds), it was found that the haplogroup distribution was most similar to Iran, and Middle Eastern countries in general, but differed significantly from countries in the Arabian Peninsula

(Al-Zahery, et al., 2003). The most prevalent haplogroups were H and U. H was found at 19.9% which is much lower than is found in Europe (30-50%) but higher than in Central Asia (14%). Haplogroup U was found in 19.0% of individuals which is similar to what is found in European populations. An additional 39.0% is represented by Western Eurasian haplogroups HV (10.6%), J (9.3%), T (8.8%), K (3.2%), X (2.8%), I (1.9%), W (1.9%), and V (0.5%). The presence of V at this low frequency is likely due to gene flow from Europe, while the presence of M (1.4%) and B (0.9%) indicate gene flow from Central Asia, and L1 (1.4%) and L2 (2.8%) represent gene flow from Africa (Al-Zahery, et al., 2003).

The mtDNA of Marsh Arabs found a significantly greater East/Southwest Asian component (11.8%) (Al-Zahery, et al., 2011. Additionally, the Asian haplotypes differed between the two with M and R2 found in the Marsh Arabs and R5a and U2d found in the general population. The two populations also differed in African haplotypes with Marsh Arabs having a much larger African component (2.8% NE African; 4.9% Sub-Saharan) than the general Iraqi Arab population that was sampled (1.2% NE African; 0.1% Sub-Saharan). Both populations had West Eurasian haplogroups, but the general population had a slightly larger European component than the Marsh Arabs (56.3% in the general population; 44.9% in Marsh Arabs) (Al-Zahery, et al., 2011).

A study of HVS-I revealed that the bulk of the existing mtDNA lineages that entered Europe from the Neat East arrived in numerous "waves" during the Upper Paleolithic and that the surviving lineages stem from either a founder effect or a bottleneck that occurred ~20,000 ya, during the last glacial maximum (LGM) (Richards, et al., 2000). They speculated that the Neolithic component comprises less than one-quarter of the modern European mtDNA gene pool and that there has been considerable back-migration into the Near East (Richards, et al., 2000).

A subsequent mtDNA study revealed female gene flow from sub-Saharan Africa into the Near Eastern gene pool (Richards, et al., 2003). This study found L1-L3A at 9%, U6 at 1%, M1 at 1%, and pre-HV at 4% in Iraqi Arabs, and found similar frequencies in Arab populations in Palestine, Jordan, and in the Bedouins. However, these lineages were either miniscule or absent in non-Arab populations (including Jews, Turks, Kurds, Armenians, Azeris, and Georgians) indicating that African gene flow was specifically into Arab populations. The authors suggest that the absence of L1-L3A in Near Eastern Jewish populations indicate that the gene flow occurred after the founding of these communities. They believe most of the gene flow postdates 5th century BC. They also found little evidence of male gene flow as there were very little sub-Saharan African Y-chromosome haplotypes in the Arab populations. Their interpretation of these findings is that the mitochondrial lineages were brought into the region via the Arab slave trade. Male slaves left relatively few descendants because they were mainly employed in manual labor and military service. Those males that were in households were employed as eunuch. Female slaves, on the other hand, were imported specifically for sex. Manumission was a regular practice and the offspring of these pairings were free individuals and became integrated into the Arab society (Richards, et al., 2003).

Autosomal

Another study used 49 autosomal SNPs to compare populations from Iraq, Turkey, Israel, Pakistan, India, China, Taiwan, Japan, Siberia, Algeria, Somalia, Uganda, Mozambique, Angola, Nigeria, Denmark, Portugal, and Spain (Tomas, Diez, Moncada, Borsting, & Morling, 2013). Using AMOVA and a step-down Holm-Sidak correction, they found that Iraq was significantly different from all these countries except for Turkey (Tomas, Diez, Moncada, Borsting, &

Morling, 2013). This finding was confirmed by another study which also found no significant difference between Iraq and Turkey (Farzad, et al., 2013).

Iran, Iraq, and Kuwait were compared using a VNTR in the DAT1 gene (Banoei, et al., 2008). Results found a high relationship probability between Turkmen in Iran and the Iraqi Arab population indicating gene flow into the Iraq Arab population from the large Turkmen population in Iraq. Comparison between the Arab populations of Iran, Iraq, and Kuwait found similarity between Iraqi and Kuwaiti Arabs while both were significantly different from Iranian Arabs which could mean that Iranian Arabs had a separate origin or the result of intermixture in the communities (Banoei, et al., 2008).

Chapter 3: Methods

Sample Collection

Samples were collected by researchers at the Forensic DNA Center for Research and Training at Al-Nahrain University in Baghdad, Iraq. Buccal swabs were taken from 1061 laboratory workers and patients at hospitals and private laboratories in Diyala (n = 139), Anbar (n = 132), Wasit (n = 120), Najaf (n = 119), Baghdad (n = 354), and Basra (n = 198).

DNA was extracted, amplified, and typed at Al-Nahrain University. Extraction was done with the PrepFiler™ Forensic DNA Extraction Kit (Applied Biosystems, Inc.), amplification was done using AmpFℓSTR® Identifiler® Kit (Applied Biosystems, Inc.), and typing was performed on an ABI PRISM® 3130XL Genetic Analyzer (Applied Biosystems, Inc.). The AmpFℓSTR® Identifiler® Kit amplifies fifteen short tandem repeats (STRs): D8S1179, D21S11, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, D2S1338, D19S433, vWa, TPOX, D18S51, D5S818, and FGA. These STRs are commonly used for forensic analyses.

The data were sent to the Laboratory of Biological Anthropology at the University of Kansas for statistical analysis. They were sent in the form of an Excel spreadsheet which contained for each individual an ID number, city of collection, and the two alleles found at each locus. The alleles were given as the number of repeats present.

Population Structure

Analyses were performed using various packages in R (R Core Team, 2016). The raw genotype data were stored in Genepop format (Raymond & Rousset, 1995; Rousset, 2008) and imported into R using the *read.genepop* function in the "adegenet" package (Jombart, 2008; Jombart & Ahmed, 2011).

Hardy-Weinberg Equilibrium

Each of the 15 loci were tested for departures from Hardy-Weinberg equilibrium in each city using the *hwx.test* function of the "HWxtest" package (Engels, 2009) in R (R Core Team, 2016). For two alleles, the HW equilibrium equation is shown in equation 1.

$$p^2 + 2pq + q^2 = 1 \quad (1)$$

where p is the frequency of one allele and q is the frequency of the other allele. This equation can be expanded for however many alleles are present at a locus (Castle, 1903; Hardy, 1908; Weinberg, 1908).

Traditionally, deviations from HW equilibrium have been tested using a χ^2 goodness-of-fit test, but this becomes unreliable when studying multi-allelic markers (Engels, 2009). STRs, which were used for this research project, are multi-allelic markers. When a χ^2 test is inappropriate, an exact test is preferred (Haldane, 1954; Levene, 1949). In this method, the potential outcomes that have the same allele frequencies as was observed are considered and the proportion of outcomes that deviate from the HW null hypothesis by at least as much as the observed outcome is the *p*-value. However, for large datasets, complete enumeration of all the tables is computationally unreasonable.

However, a publication showed that the set of contingency tables needed for Fisher's exact test could be represented as a network of nodes connected by arcs traversing from the initial node to the final node (Mehta & Patel, 1983). Each arc represents one of the contingency tables and the length of that arc is the probability. Rather than calculating the probability for each arc, a confidence interval could be generated by calculating the probability for the shortest and longest arcs (Mehta & Patel, 1983).

This network algorithm, which was for square tables, was adapted for triangular tables of genotype data in which each table with a given set of allele counts is an arc in the network and each node is the residual allele counts that that point (Engels, 2009). The triangular matrix of genotype frequencies is as follows in equation 2:

$$\begin{bmatrix} a_{11} & & & \\ a_{21} & a_{22} & & \\ \vdots & \vdots & \ddots & \\ a_{k1} & a_{k2} & \dots & a_{kk} \end{bmatrix} \quad (2)$$

Each field in the matrix is represented by a value, a_{ij}, which is the observed number of genotypes with alleles *i* and *j*. The test statistics that are calculated include the likelihood ratio (*LR*) and the conditional probability (*P*) which are given by equations 3 and 4:

$$LR(a) = \frac{\prod_i m_i^{m_i}}{2^{n+d} n^n \prod_{i \geq j} a_{ij}^{a_{ij}}} \quad (3)$$

$$P(a|m) = \frac{2^{n-d} n! \prod m_i!}{(2n)! \prod_{i \geq j} a_{ij}!} \quad (4)$$

where m_i is the number of *i* alleles, *n* is the total sample size, and *d* is the number of homozygotes. Test statistics are calculated for each table and included in the recursive algorithm which traverses the entire network of tables, computing the probabilities and test statistics for each table, and providing an exact *p*-value (Engels, 2009).

However, computing all the tables for large datasets with multiple alleles can be too computationally intensive (Engels, 2009). For example, a dataset containing only five alleles with counts of m = [9 6 3 1 1] has 139 tables that need to be calculated for the exact *p*-value. In cases of large datasets, a Monte Carlo test is preferred (Guo & Thompson, 1992). The *hwx.test* function of the "HWxtest" package first determines the number of tables needed for a complete enumeration and if that exceeds 1 x 10^{10} tables then it performs a Monte Carlo test instead. In this test, a specified number of random tables of genotypes of the observed allele counts are

generated, the probability and test statistics are computed for each random table, and an estimate of the *p*-value is generated (Engels, 2009). The Monte Carlo method was employed for this dataset using 1 x 10^5 permutations.

Random permutations for the Monte Carlo procedure were generated using the Fisher-Yates shuffle (Fisher & Yates, 1943) which is performed by selecting a random number, *k,* between one and the total number needed and then the *k*th element is selected. This procedure is then repeated, without replacement, until all the original elements have been shuffled into a random sequence. The random numbers are generated using the multiply-with-carry method (Marsaglia, 2003) in which $t_n = (ax + c)$. The next *x* in the sequence is determined by $x = t \ mod \ b$, where *mod b* is modulo base *b*, usually $b = 2^{32}$; the next *c* in the sequence is determined by $c = \frac{t}{b}$, rounded to the nearest integer.

The output from the *hwx.test* (Engels, 2009) includes *p*-values generated from the log likelihood ratio (*LLR*), conditional probability, *U*-score (Rousset & Raymond, 1995), and χ^2. Engels (2009) recommends using the *LLR p*-value when there is no expectation for which direction the data might deviate from HW equilibrium. As this was the case for these data, the reported *p*-values are based on *LLR*. These *p*-values were then corrected for multiple testing using the Holm-Bonferroni method (Holm, 1979). In this method, the tests are ranked according to their *p*-values with the smallest *p*-value having a rank of one. Each test's *p*-value is then sequentially compared to a calculated threshold to determine whether it should be considered significant. The *p*-value of a test is considered significant if:

$$p_k < \frac{\alpha}{m+1-k} \tag{5}$$

where α is the overall significance level, *m* is the total number of tests, and *k* is the rank number of the test (Holm, 1979).

F-Statistics

F-statistics, also known as fixation indices (Wright, 1943), were calculated using the "pegas" package (Paradis, 2010). The *as.loci* function of "pegas" was used to convert the dataset from "adegenet" format (genind object) where individuals are in rows and alleles are in columns to the "pegas" format (loci object) which also has individuals as rows but has loci in the columns with the alleles in the data file separated by a forward slash. The *Fst* function was then used to calculate the F_{IS}, F_{ST}, and F_{IT} values at each locus for the entire sample set. This function uses formulae provided by Weir and Cockerham (1984) and Weir and Hill (2002):

$$a = \frac{\bar{n}}{n_c}\left\{s^2 - \frac{1}{\bar{n}-1}\left[\bar{p}(1-\bar{p}) - \frac{r-1}{r}s^2 - \frac{1}{4}\bar{h}\right]\right\} \quad (6)$$

$$b = \frac{\bar{n}}{\bar{n}-1}\left[\bar{p}(1-\bar{p}) - \frac{r-1}{r}s^2 - \frac{2\bar{n}-1}{4\bar{n}}\right] \quad (7)$$

$$c = \frac{1}{2}\bar{h} \quad (8)$$

in which

$\bar{n} = \sum \frac{n_i}{r_i}$, the average sample size (9)

$n_c = \frac{r\bar{n} - \sum \frac{n_i^2}{r\bar{n}}}{r-1}$ (10)

$\bar{p} = \sum \frac{n_i \tilde{p}_i}{r\bar{n}}$, the average sample frequency of the allele (11)

$s^2 = \sum \frac{n_i(\tilde{p}_i - \bar{p})^2}{(r-1)\bar{n}}$, the sample variance of the allele frequency over populations (12)

$\bar{h} = \sum \frac{n_i \tilde{h}_i}{r\bar{n}}$, the average heterozygote frequency of the allele (13)

where $\tilde{p}_i$ is the frequency of the allele in a population, *i*, of size n_i; *r* is the total number of populations; and $\tilde{h}_i$ is the observed proportion of heterozygotes for the allele in population *i*.

F_{IS}, the inbreeding coefficient, is calculated by:

$$F_{IS} = 1 - \frac{c}{b+c} \quad (14)$$

F_{ST}, the fixation index, is calculated by:

$$F_{ST} = \frac{a}{a+b+c} \quad (15)$$

F_{IT}, the overall fixation index, is calculated by:

$$F_{IT} = 1 - \frac{c}{a+b+c} \quad (16)$$

The "pegas" package provides these values for each locus across all subpopulations, and overall values were obtained by averaging across all loci.

Nei's Genetic Distance

Nei's genetic distance was calculated using the *dist.genpop* function of the "adegenet" package (Jombart, 2008; Jombart & Ahmed, 2011). The *genind2genpop* function was used to convert the dataset from genotype data (genind object) into allele counts per population (genpop object), which is the format needed to use the *dist.genpop* function. Method 1 (Nei's distance) was selected for the function.

The first step in obtaining Nei's genetic distance values is to calculate Nei's genetic identity (*I*) for an allele at a locus:

$$I = \frac{\sum_{i=1}^{k} p_{ix} p_{iy}}{\sqrt{\sum_{i=1}^{k} p_{ix}^2 \sum_{i=1}^{k} p_{iy}^2}} \quad (17)$$

where p_{ix} is the frequency of allele *i* in population *x*, and p_{iy} is the frequency of allele *i* in population *y*. The second step is to sum these values over all loci and alleles, and divide by the total number of loci. Nei's genetic identity ranges from 0 to 1. The final step is to use *I* to calculate *D*, Nei's genetic distance, using $D = -\ln(I)$ (Nei, 1972; Nei, 1978).

Principal Component Analysis

Principal component analysis (Pearson, 1901) was performed using the "ade4" package (Chessel, Dufour, & Thioulouse, 2004; Dray & Dufour, 2007; Dray, Dufour, & Chessel, 2007) and the "adegenet" package (Jombart, 2008; Jombart & Ahmed, 2011). A principal component analysis (PCA) was done using the individual counts. First, the *scaleGen* function of the "adegenet" package was used to compute scaled allele frequencies from the raw data and place them into a matrix format. Then the *class* function of the "base" package (R Core Team, 2016) was used to confirm that the dataset was a matrix and the *dim* function from the same package was used to check that the dimensions of the matrix were correct. Then, the principal component analysis (PCA) was done using the *dudi.pca* function of the "ade4" package. The eigenvalues were graphed using the *barplot* function of the "graphics" package (R Core Team, 2016). The PCA were plotted using the *s.label* function of the "aed4" package. A second PCA was done with two individuals removed from the dataset (explained further in Chapter 5: Discussion). It was plotted using the *s.class* function of the "ade4" package.

PCAs were also performed comparing the allele frequencies of the cities to each other and comparing the allele frequencies of Iraq to surrounding countries. Frequency data were taken from previous studies done on Turkey (Yavuz & Sarikaya, 2005), Syria (Abdin, Shimada, Brinkmann, & Hohoff, 2003), Saudi Arabia (Osman, et al., 2015), Kuwait (Alenizi, Goodwin, Ismael, & Hadi, 2008), and Iran (Shepard & Herrera, 2006). These data were stored in Excel files and imported into R using the *read.xlsx* function of the "openxlsx" package (Walker, 2015). The PCAs were done using the *dudi.pca* function of "ade4" and plotted using the *colorplot* function of the "adegenet" package for PCAs.

PCA organizes the data into a matrix and uses matrix algebra to reduce the variance into eigenvalues and eigenvectors using the following steps:

1. Organize the *n* samples of *m*-dimensional data as vectors, $\vec{x}_1, \ldots, \vec{x}_n$.

2. Take the sample mean of the variables, $\vec{Y} = \frac{\vec{x}_1 + \cdots + \vec{x}_n}{n}$, and subtract the sample mean from each sample vector and place them in a *m x n* matrix, *B*:

$$\vec{x}_1 = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \quad \vec{x}_2 = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \quad \vec{x}_n = \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_n \end{bmatrix} \quad \vec{Y} = \begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 \\ \vdots \\ \bar{Y}_n \end{bmatrix} \tag{18}$$

and

$$B = \begin{bmatrix} a_1 - \bar{Y}_1 & b_1 - \bar{Y}_1 & \ldots & n_1 - \bar{Y}_1 \\ a_2 - \bar{Y}_2 & b_2 - \bar{Y}_2 & \ldots & n_2 - \bar{Y}_2 \\ \vdots & \vdots & \ldots & \vdots \\ a_n - \bar{Y}_n & b_n - \bar{Y}_n & \ldots & n_n - \bar{Y}_n \end{bmatrix} \tag{19}$$

3. Then create the *m x m* covariance matrix (*S*) so that S_{ii} is the variance of the *i*th variable and S_{ij} is the covariance of the *i*th and *j*th variables. For example:

$$S_{11} = \frac{((a_1 - \bar{Y}_1)^2 + (b_1 - \bar{Y}_1)^2 + \cdots + (n_1 - \bar{Y}_1)^2)}{n-1} \tag{20}$$

is the variance of the first variable, and

$$S_{21} = \frac{(a_1 - \bar{Y}_1)(a_2 - \bar{Y}_2) + (b_1 - \bar{Y}_1)(b_2 - \bar{Y}_2) + \cdots + (n_1 - \bar{Y}_1)(n_2 - \bar{Y}_2)}{n-1} \tag{21}$$

is the covariance of the first and second variables.

4. Then, find the eigenvalues (λ_i) and the orthogonal set of eigenvectors ($\vec{v}_i$) for the *m x m* matrix, *S*. The eigenvalues are found according to the following formula:

$$|S - \lambda I| = 0 \tag{22}$$

where *S* is the *m x m* matrix and *I* is an identity matrix.

For example, for a 2 x 2 matrix the eigenvalues would be calculated as:

$$|S - \lambda I| = \left|\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right| = \left|\begin{bmatrix} S_{11} - \lambda & S_{12} \\ S_{21} & S_{22} - \lambda \end{bmatrix}\right|$$

$$= |(S_{11} - \lambda)(S_{22} - \lambda) - (S_{12})(S_{21})| = 0$$

This will give a quadratic equation for which there are two possible values of λ. They are arranged in decreasing order with the largest being λ_1, and so on. The eigenvalues are then used to find the eigenvectors from the following formula:

$$(S - \lambda_i)\vec{v}_i = 0 \quad (23)$$

The direction given by $\vec{v}_1$ (the first principal component) accounts for the fraction of the total variance *T* explained by λ_1, so the vector points in the most significant direction (Hotelling, 1933a; Hotelling, 1933b). And among all the directions that are orthogonal (i.e. perpendicular) to $\vec{v}_1$, $\vec{v}_2$ (the second principal component) points in the most significant direction. It is not uncommon for most of the variation in the data to be explained among the first few principal components, which allows for the dataset to be reduced from several dimensions to just two or three which can be easily visualized in a plot.

Analysis of Molecular Variance

The *poppr.amova* function of the "poppr" package was used to perform an analysis of molecular variance for this data (Kamvar, Tabima, & Grünwald, 2014; Kamvar, Brooks, & Grünwald, 2015). The *strata* function from "adegenet" defined the population stratification of the dataset while still in "adegenet" format (genind object). The *as.genclone* function of "poppr" then converted the data from "adegenet" format (genind object) to "poppr" format (genclone object) which includes a `mlg` slot which retains multilocus genotypes regardless of how the data is subset. The *strata* output was then visualized using "base" package's *table* function (R Core Team, 2016) to ensure that it was properly stratified. The *poppr.amova* function of "poppr" acts as a wrapper for the *amova* function of "ade4" (Chessel, Dufour, & Thioulouse, 2004; Dray &

Dufour, 2007; Dray, Dufour, & Chessel, 2007). It automates all the elements needed for the "ade4" function by generating a distance matrix of all unique genotypes in the dataset, constructing a data frame defining the hierarchy of the distance matrix (based on the output of the *strata* function), and putting together a genotype frequency table. After *poppr.amova* constructs all of the elements needed, then it utilizes the *amova* function of the "ade4" package to perform the AMOVA.

Analysis of molecular variance (AMOVA) is a statistical method developed to detect population differentiation using molecular markers (Excoffier, Smouse, & Quattro, 1992). It was originally designed for work with DNA haplotypes, but can be applied to any marker system. The first step in AMOVA is to calculate a pairwise *N x N* genotypic distance matrix. The matrix is organized so that individuals in a pairwise fashion at each locus and assigned a value ranging from 0 to 4, according to the following rules:

1. 0 is assigned if the individuals share a genotype (Ex. *ii* and *ii*, or *ij* and *ij*)
2. 1 is assigned if they differ by a single allele (Ex. *ii* and *ij*, or *ij* and *ik*)
3. 2 is assigned if they share no alleles and both are heterozygous (Ex. *ij* and *kl*)
4. 3 is assigned if the individuals share no alleles and one is homozygous (Ex. *ii* and *jk*)
5. 4 assigned if they share no alleles and both are homozygous (Ex. *ii* and *jj*)

This is a Euclidean squared distance metric: $d^2(genotype\ 1, genotype\ 2)$. The distances are then summed across the loci, assuming that the loci are independent (Excoffier, Smouse, & Quattro, 1992).

The sum of squares (*SS*) and mean sums of squares (*MS*) are then calculated from the distance matrix:

$$SS_{TOT} = \frac{\sum d_{ij}^2}{2N} \tag{24}$$

$$SS_{WP1} = \frac{\sum d_{ij}^2}{2n_1} \quad (25)$$

$$SS_{WP} = SS_{WP1} + SS_{WP2} + \cdots + SS_{WPn} \quad (26)$$

$$SS_{AP} = SS_{TOT} - SS_{WP} \quad (27)$$

$$MS_{WP} = \frac{SS_{WP}}{df_{WP}} \quad (28)$$

$$MS_{AP} = \frac{SS_{AP}}{df_{AP}} \quad (29)$$

where d_{ij}^2 = the squared genetic distance between the *i*th and *j*th sample, df_{AP} = the number of populations – 1, and df_{WP} = the total number of samples – the number of populations (Excoffier, Smouse, & Quattro, 1992).

Variance estimates are calculated by:

$$s_{WP}^2 = MS_{WP} \quad (30)$$

$$s_{AP}^2 = \frac{MS_{among} - MS_{within}}{N_0} \quad (31)$$

$$N_0 = \frac{1}{(a-1)} \times \left(\sum_{k=1}^{a} n_k - \left(\frac{\sum_{k=1}^{a} n_k^2}{\sum_{k=1}^{a} n_k} \right) \right) \quad (32)$$

where a = the number of populations and n_k = the number of samples in the *k*th population (Excoffier, Smouse, & Quattro, 1992).

The within-individual variation is calculated by comparing the alleles at each locus within an individual and assigning a 0 if the individual is homozygous at that locus and a 1 if the individual is heterozygous at that locus (Excoffier, Smouse, & Quattro, 1992). These values are then summed across all loci and all individuals and then divided by 2 to provide the within individual sum of squares. Degrees of freedom are the total number of individuals and the within individual mean square is the within individual sum of squares divided by its degrees of freedom (Excoffier, Smouse, & Quattro, 1992).

ϕ-statistics are statistics calculated with AMOVA that follow the structure of *F*-statistics in order to characterize variation at different hierarchical levels within the population (Excoffier, Smouse, & Quattro, 1992). ϕ_{ST} describes the variance within random genotypes (within individual variance) relative to variance among random genotypes taken from the entire population (total variance). Given that:

σ_a^2 = variance between populations

σ_b^2 = variance between samples

σ_c^2 = variance within samples

σ^2 = total variance

then, ϕ_{ST} is calculated as:

$$\phi_{ST} = \frac{\sigma_a^2+\sigma_b^2}{\sigma^2} \quad (33)$$

The smaller the value of ϕ_{ST}, the greater the within individual variance, indicating less population substructure (Excoffier, Smouse, & Quattro, 1992).

ϕ_{CT} describes the variance in random genotypes within the designated subpopulations (between-populations variance) relative to variance among random genotypes taken from the entire population (total variance) (Excoffier, Smouse, & Quattro, 1992). It is calculated by:

$$\phi_{CT} = \frac{\sigma_a^2}{\sigma^2} \quad (34)$$

Larger values of ϕ_{CT} indicate greater between-population variation and so, unlike ϕ_{ST}, a larger value for ϕ_{CT} suggests genetic differentiation between the designated subpopulations (for this data, cities).

ϕ_{SC} measures the variance among random genotypes within the population (between-samples variance) relative to that of random pairs of genotypes drawn from within the subpopulations (total variance without the between populations variance). It is calculated by:

$$\phi_{SC} = \frac{\sigma_b^2}{\sigma_b^2+\sigma_c^2} \quad (35)$$

Larger values of ϕ_{SC} indicate greater variation between samples. Looking at these three statistics provides information as to which level within the population contains the greatest variance and can be evaluated for evidence for genetic differentiation between subpopulations (Excoffier, Smouse, & Quattro, 1992).

To determine whether the variance components and ϕ-statistics estimated at each level can be considered statistically significant, a *p*-value must be obtained. Excoffier et al (1992) recommend using a randomization test for this. The null hypothesis under which the test is performed is that all variation in the population is within samples and no variation is between samples or between populations. This dataset was too large to do a complete enumeration of all possible outcomes, so a sampled randomization test was done. In a sampled randomization test, all the data is re-sorted randomly without replacement and the test statistics are calculated from the randomized data. This is repeated for a specified number of times and the distribution of the estimates obtained from all the randomized permutations form the null distributions for each estimate. Then, the proportion of times that the observed estimate occurs in null distribution is given as the *p*-value for that estimate. The *randtest* function of the "ade4" package was used to perform the sampled randomization test (Chessel, Dufour, & Thioulouse, 2004; Dray & Dufour, 2007; Dray, Dufour, & Chessel, 2007). The *set.seed* function from the "base" package (R Core Team, 2016) was used to set a seed for the randomization test which used 999 permutations to obtain *p*-values. The *plot* function of the "graphics" package (R Core Team, 2016) was used to plot the output of the randomization test.

Discriminant Analysis of Principal Components

Discriminant analysis of principal components (DAPC) is a multivariate method to identify and describe genetic clusters (Jombart, Devillard, & Balloux, 2010). This method involves performing a discriminant analysis (DA) on data that has been transformed into principal components (PCs) rather than the original dataset. The rationale behind the development of this method is that it allows DA to be performed on genetic data which often violates its assumptions. DA requires that the number of alleles be less than the number of individuals sampled and that there is no correlation between allele frequencies (Lachenbruch & Goldstein, 1979). Transforming the data using PCA ensures that the number of variables (PCs) is less than that of the sample individuals and that the variables are completely uncorrelated (Jombart, Devillard, & Balloux, 2010).

Two DAPCs were performed on this dataset: the first DAPC was performed on data sorted by inferred genetic clusters obtained using a *k*-means algorithm and the second DAPC was performed on data sorted by city. They were performed using the "adegenet" package (Jombart & Ahmed, 2011; Jombart, 2008). The *find.clusters* function was used to sort the data into genetic clusters. First, it transforms the data using PCA retaining the number of PCs specified by the user. Second, it runs the transformed data through the *k*-means clustering algorithm multiple times for increasing values of *k* (the total number to be done is specified by the user). Third, it provides the Bayesian information criterion (BIC) for each value of *k*, and the user selects the number of clusters to use based on this information.

The *find.clusters* function utilizes "stats" package's *kmeans* function (R Core Team, 2016) which uses the *k*-means clustering algorithm detailed in Hartigan and Wong (1979). The method defines *k* centers for each cluster, placing the centers as far as possible from each other.

Each point in the dataset is assigned to the nearest center. Once all points are assigned to a center, *k* new centroids are calculated at the barycenters of the clusters that resulted from the previous steps. After these new centroids have been established, it goes back to the beginning and each data point is assigned to one of the new centroids. This process loops, with the *k* centers changing their location with each loop. This repeats until the centers do not move any more (Hartigan & Wong, 1979).

Bayesian information criterion (BIC) was used to select the best fit for the number of genetic clusters. The BIC helps to protect against model overfitting due to selecting too many parameters by penalizing the number of parameters in the model (Schwarz, 1978). For genetic data, it is calculated as follows:

$$BIC = n \log (W(X)) + g \log (n) \quad (36)$$

where *(W(X))* is the residual variance (within group variance) and *g* is the number of groups. BIC quantifies how well the model fits the data and penalizes use of too many clusters. The optimal number of clusters has the lowest BIC (Schwarz, 1978). The *table* function from the package "base" (R Core Team, 2016) was used to visualize the *find.clusters*'s output.

A χ^2 test of independence (Pearson, 1900) was performed to test whether an individual's city was independent of the genetic cluster they were assigned to. The data was stored in a contingency table and expected values were calculated by treating it as a pure contingency table, $E_i = \frac{row\ total \times column\ total}{N}$, and the test-statistic was calculated by:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (37)$$

where O_i is the observed value and E_i is the expected value. The test-statistic was then compared to a χ^2 distribution to obtain a *p*-value.

Then, the *dapc* function was used to perform a discriminant analysis on the transformed data which had been sorted into groups by genetic cluster. DA is similar to PCA in that they both look for linear combinations of variables (principal components in PCA and discriminant functions in DA) to explain the data. However, PCA summarizes the total variability while DA summaries the between-group variability (Jombart, Devillard, & Balloux, 2010). DAPC uses linear discriminant analysis (LDA) which was formulated by Ronald Fisher (1936) for discriminating between two groups and was later expanded for multiple groups by CR Rao (1948). DA is performed using the following steps (Venables & Ripley, 2002; Jombart, Devillard, & Balloux, 2010):

1. Compute two *g x g*-dimensional matrices, a within-group and a between-group scatter matrix. Within-group scatter matrix, S_W, is calculated by:

$$S_W = \sum_{i=1}^{g} S_i \tag{38}$$

where $S_i = \sum_{x \in D_i}^{n} (x - m_i)(x - m_i)^T$ and $m_i = \frac{1}{n_i} \sum_{x \in D_i}^{n} x_k$. Between-group scatter matrix, S_B, is calculated by:

$$S_b = \sum_{i=1}^{g} N_i (m_i - m)(m_i - m)^T \tag{39}$$

where m is the overall mean, m_i is the sample mean of the group, and N_i is the sample size of the group.

2. Find the eigenvectors and eigenvalues:

$$Av = \lambda v \tag{40}$$

where $A = S_W^{-1} S_B$, v = eigenvector, and λ = eigenvalue

3. Sort the eigenvectors by their eigenvalues from highest to lowest.
4. Construct a *k x d*-dimensional eigenvector matrix, *W*. (At this point, the *dapc* function provides the user with a graph of the discriminant functions (eigenvectors) showing their

F-statistics and the user can select how many to keep; if the user does not select to keep them all then the function will drop the ones with the lowest eigenvalues).

5. Transform the samples onto the new subspace:

$$Y = X \times W \tag{41}$$

where X is a $n \times d$-dimensional matrix representing the n samples.

A DAPC was then performed on the dataset using the cities as the *a priori* groupings. To select the appropriate number of PCs to retain for the DA, the "adegenet" package offers two options: a-score optimization and DAPC cross-validation (Jombart, 2008; Jombart, Devillard, & Balloux, 2010; Jombart & Ahmed, 2011). It is important to select an appropriate number of PCs because retaining too few will result in loss of necessary information while retaining too many can lead to over-fitting of the model.

The a-score is the difference between the proportion of successful reassignment (observed discrimination) and a value obtained using random groups (random discrimination) (Jombart, 2008; Jombart, Devillard, & Balloux, 2010; Jombart & Ahmed, 2011). Essentially, it is the percentage of correct assignment corrected for the number of PCs that have been retained. With the *optim.a.score* function in "adegenet", the user specifies the number of PCs to retain and the function repeats the DAPC analysis using randomized groups and computes the a-score for a certain number of the retained PCs (ex. every 5, every 10, etc. depending on how many PCs the user has chosen to retain) and then does a spline interpolation to estimate the optimal number of PCs to retain. However, this optimization is impacted by the number of PCs the user chooses to retain because the number of a-scores and which number of PCs a-scores will be calculated before spline interpolation will change depending on the number of PCs the function is told to

retain. Therefore, this function should not be used by itself, but in conjunction with other methods (Jombart, 2008; Jombart, Devillard, & Balloux, 2010; Jombart & Ahmed, 2011).

For the DAPC cross-validation, "adegenet" package's *as.matrix.genind* (Jombart, 2008; Jombart & Ahmed, 2011) was used to convert the original dataset from genind format into a matrix, then *pop* was used on the original dataset to output what population each individual belongs to. Both of these outputs were then put into the *xvalDapc* function which used the *as.matrix.genind* output to perform the analysis while referencing the *pop* output for the group membership of the individuals. Cross-validation involves dividing the data into a training set (comprising 90% of the data, by default) and a validation set (comprising the remaining 10% of the data, by default) (Jombart, 2008; Jombart, Devillard, & Balloux, 2010; Jombart & Ahmed, 2011). Selection of individuals for the sets is done with stratified random sampling to ensure that at least member of each group ends up in both the training and validation sets. DAPC is then carried out on the training set for a certain number of PCs and is then used to assign individuals from the validation set. The percent of correct assignment for the validation set is then calculated. This is replicated a certain number of time (default 30) for each level of PC retention. A graph is then generated with the number of PCA axes retained on the x-axis and the proportion of successful outcome prediction on the y-axis. Predictive success is sub-optimal when there are both too few and too many PCA axes retained. The optimal number of PCs is the one with the highest mean success and the lower root mean squared error. While these may be the same, they may not be. In that case, the authors recommend that the number of PCs with the lowest root mean squared error should be chosen (Jombart, 2008; Jombart, Devillard, & Balloux, 2010; Jombart & Ahmed, 2011).

Multidimensional Scaling

Multidimensional scaling (MDS) is a method for visualizing the level of similarity between groups (Torgerson, 1952). It is similar to PCA in that both methods seek to reduce the dataset down into the fewest number of dimensions that retain the variability in the data, but PCA does so using a centered covariance matrix while MDS uses a double-centered distance matrix (Borg & Groenen, 1997).

Allele frequency data were collected for Turkey (Yavuz & Sarikaya, 2005), Syria (Abdin, Shimada, Brinkmann, & Hohoff, 2003), Saudi Arabia (Osman, et al., 2015), Kuwait (Alenizi, Goodwin, Ismael, & Hadi, 2008), Iran (Shepard & Herrera, 2006), Poland (Szczerkowska, Kapińska, Wycocka, & Cybulska, 2004), Belgium (Decorte, Gilissen, & Cassiman, 2003), Japan (Hashiyada, 2000), China (Wang, Yu, Wang, Li, & Jin, 2005), Angola (Beleza, et al., 2005), and Equatorial Guinea (Alves, et al., 2005).

The frequency data were stored in an Excel file and *read.xlsx* from "openxlsx" package (Walker, 2015) were used to input the file into R. The *dist* function from "stats" (R Core Team, 2016) was then used to compute the Euclidean distances between the countries and return it in matrix format. This matrix was then inputted into the *cmdscale* function from "stats" to perform classical (metric) multidimensional scaling. The *plot* function from "graphics" (R Core Team, 2016) was used to visualize the MDS output in a 2D plot. The *scatter3D* function of the "plot3D" package was used to make a 3D plot (Soetaert, 2017).

The Euclidean distance between two groups *i* and *j* is calculated by:

$$d_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} \tag{42}$$

where *x* and *y* represent coordinates (in this dataset, allele frequencies). The Euclidean distances are then squared to form a squared distance matrix, *D*. Classical MDS is performed using the following steps:

1. Set up the matrix of squared distances, *D*.
2. Put matrix *D* through the double centering equation to obtain matrix *B*:

$$B = -\frac{1}{2}JPJ \tag{43}$$

where *J* is a matrix described as $J = I - \frac{1}{n}11'$, where *I* is an *n x n* identity matrix, *n* is the number of groups, and $11'$ an all-ones matrix.

3. Extract the *m* largest eigenvalues of *B* and the corresponding eigenvectors.
4. The *m*-dimensional spatial configuration of the groups is derived from the coordinate matrix, *X*, which is found by:

$$X = E_m\sqrt{\Lambda_m} \tag{44}$$

where E_m is the matrix of *m* eigenvectors and Λ_m is the diagonal matrix of *m* eigenvalues.

These coordinates can then be used to graphically display the differences between the groups.

Forensic Applications

For the establishment of a forensic DNA database for Iraq, PowerStats v1.2 (Tereba, 1999) was used to calculate allele frequencies, heterozygosities, homozygosities, polymorphism information content, matching probabilities, power of discrimination, power of exclusion, and the typical paternity index at each locus for each city. Allele frequencies were calculated by counting the number of time that allele occurred and dividing it by the total number of alleles. Heterozygosity was determined by counting the number of heterozygous individuals and dividing it by the total of number of individuals. Homozygosity was calculated by counting the number of homozygous individuals and dividing it by the total number of individuals.

Polymorphism information content (PIC) is a measure of a locus's usefulness in differentiating people (Botstein, White, Skolnick, & Davis, 1980). A marker is highly informative if a randomly chosen individual is highly likely to be heterozygous at that marker. The larger a value of PIC the marker has, more useful it is in forensics. PIC is calculated by:

$$PIC = 1 - \sum_{i=1}^{n} p_i^2 - (\sum_{i=1}^{n} p_i^2)^2 + \sum_{i=1}^{n} p_i^4$$

where p_i is the frequency of a marker allele and n is the number of different alleles at that marker (Botstein, White, Skolnick, & Davis, 1980).

Matching probability (MP) is the probability that two randomly selected individuals will have the same genotype (Brenner & Morris, 1990). To determine the MP for a locus, the following formula was used:

$$MP = \sum_{i=a}^{n} \sum_{j \geq 1}^{n} P_{ij}^2$$

where i and j are the frequencies of all possible alleles a to n and P_{ij} are the frequencies of all the possible genotypes. MP is also expressed as $\frac{1}{MP}$, which is the number of people that would need to be sampled before finding a matching profile. So, the lower the MP of a locus, the more valuable it is for forensic purposes. The overall MP across all loci is the product of each locus's MP, assuming that these loci are independent. The power of discrimination (PD), the potential power of a locus to differentiate between any two randomly chosen individuals, is simply one minus the MP (Brenner & Morris, 1990).

For paternity testing, the power of exclusion and typical paternity index of a locus indicate how useful that marker is for excluding unrelated individuals and giving a likelihood that a matching profile is the biological father. Power of exclusion (PE) is the proportion of people that will have a DNA profile different from that of a randomly selected individual

(Brenner & Morris, 1990). That is, the power of a marker to exclude unrelated individuals. It is calculated by:

$$PE = h^2(1 - 2hH^2)$$

where *h* is the heterozygosity at that locus and *H* is the homozygosity at that locus. The total PE across all loci is found by:

$$TPE = 1 - \prod(1 - PE_i).$$

The typical paternity index (TPI) is how many times more likely it is that a matching genotype at that locus belongs to the biological father than to a randomly selected individual (Brenner & Morris, 1990). It is calculated by:

$$TPI = \frac{1}{2H}$$

where *H* is the homozygosity of that locus. The total TPI across all loci is the product of each locus's TPI, assuming that these loci are independent (Brenner & Morris, 1990).

Chapter 4: Results

Population Structure

Hardy-Weinberg Equilibrium

Departures from Hardy-Weinberg equilibrium were tested using the *hwx.test* function in the "HWxtest" package (Engels, 2009). Since the *hwx.test* function uses a Monte Carlo method (using 1 x 10^5 trials), a seed was set using the *set.seed* function from the "base" package (R Core Team, 2016). The log likelihood ratio (*LLR*) *p*-values were taken from the output and put into a table (Table 1). To correct for multiple testing, a Holm-Bonferroni procedure was performed; using this method, none of the *p*-values in the table could be considered significant.

Locus	Diyala	Anbar	Wasit	Najaf	Baghdad	Basra
D8S1179	0.978350	0.945340	0.695530	0.077240	0.152120	0.562390
D21S11	0.194890	0.068920	0.054980	0.497430	0.025690	0.380920
D7S820	0.728410	0.701240	0.555720	0.910650	0.193200	0.874440
CSF1PO	0.090950	0.663010	0.770710	0.796680	0.166190	0.359600
D3S1358	0.044130	0.957720	0.300730	0.373440	0.070350	0.072990
TH01	0.842440	0.142830	0.093800	0.005320	0.184920	0.610110
D13S317	0.472890	0.401180	0.172090	0.253100	0.081970	0.844380
D16S539	0.014190	0.194560	0.214600	0.836450	0.036610	0.178370
D2S1338	0.033080	0.648270	0.986390	0.372260	0.001870	0.076540
D19S433	0.452430	0.401120	0.772300	0.549010	0.687170	0.646600
vWA	0.226620	0.532940	0.935290	0.651600	0.114050	0.392550
TPOX	0.319710	0.126040	0.079970	0.786070	0.019860	0.525200
D18S51	0.104180	0.825380	0.391510	0.567420	0.022350	0.870120
D5S818	0.998160	0.089140	0.507940	0.533620	0.801390	0.126220
FGA	0.267860	0.826730	0.480950	0.043660	0.013510	0.150030

Table 1. *p*-Values for Hardy-Weinberg test for each city at each locus.

F-Statistics

F-statistics were calculated using the *Fst* function of the "pegas" package (Paradis, 2010). This function provides the F_{IS}, F_{ST}, and F_{IT} values for each locus in the sample set (Table 2). Negative values are allowable in Weir & Cockerham's (1984) formula and can be taken as zero,

indicating no genetic differentiation. These values were then averaged across all loci to obtain overall values of F_{IT} = 0.024746637, F_{ST} = 0.001623177, and F_{IS} = 0.023157233.

Locus	F_{IT}	F_{ST}	F_{IS}
D8S1179	0.01737203	-0.0009187090	0.018273949
D21S11	0.01421670	0.0008722232	0.013356131
D7S820	0.02025142	0.0013489352	0.018928015
CSF1PO	0.02114328	0.0022267083	0.018958784
D3S1358	0.02105124	0.0002064204	0.020849127
TH01	0.03640531	0.0013492448	0.035103430
D13S317	-0.00205339	0.0017545895	-0.003814673
D16S539	0.03035475	0.0058539084	0.024645112
D2S1338	0.04886340	0.0022652707	0.046703924
D19S433	0.02772095	0.0016085316	0.026154492
vWA	0.05512245	0.0029042921	0.052370252
TPOX	-0.01632065	0.0032671385	-0.019651997
D18S51	0.05522620	-0.0004512938	0.055652383
D5S818	0.01347378	0.0008200684	0.012664096
FGA	0.02837209	0.0012403217	0.027165463
Average	0.024746637	0.001623177	0.023157233

Table 2. *F*-statistics for collected samples at each locus.

Nei's Genetic Distance

Nei's genetic distances between each pair of cities were calculated using the *dist.genpop* function of the "adegenet" package (Jombart, 2008; Jombart & Ahmed, 2011). According to this output (Table 3), Anbar is the city with the greatest genetic distance from the other cities with Anbar-Wasit being the highest, followed by Anbar-Diyala and then Anbar-Najaf. Baghdad had the lowest genetic distance from the other cities with Baghdad-Basra being the lowest, followed by Baghdad-Diyala and then Baghdad-Najaf.

	Diyala	Anbar	Wasit	Najaf	Baghdad
Anbar	0.02835314				
Wasit	0.01596568	0.02948430			
Najaf	0.01833119	0.02749541	0.01960896		
Baghdad	0.01374901	0.01797089	0.01772762	0.01395988	
Basra	0.01846977	0.01802484	0.02319713	0.01867548	0.01353695

Table 3. Nei's genetic distance between pairs of cities sampled.

Principal Component Analysis

Principal component analyses (PCAs) were performed using the *dudi.pca* function of the "ade4" package (Chessel, Dufour, & Thioulouse, 2004; Dray & Dufour, 2007; Dray, Dufour, & Chessel, 2007) and plotted using various functions in the "ade4" and "adegenet" (Jombart, 2008; Jombart & Ahmed, 2011) packages. A PCA was done on the individual counts. The eigenvalues for the first three principal components (PCs) were: PC1 = 9.157, PC2 = 8.266, and PC3 = 2.613 (Figure 2). Since most of the variation was found in the first two PCs, a 2D plot was considered appropriate.

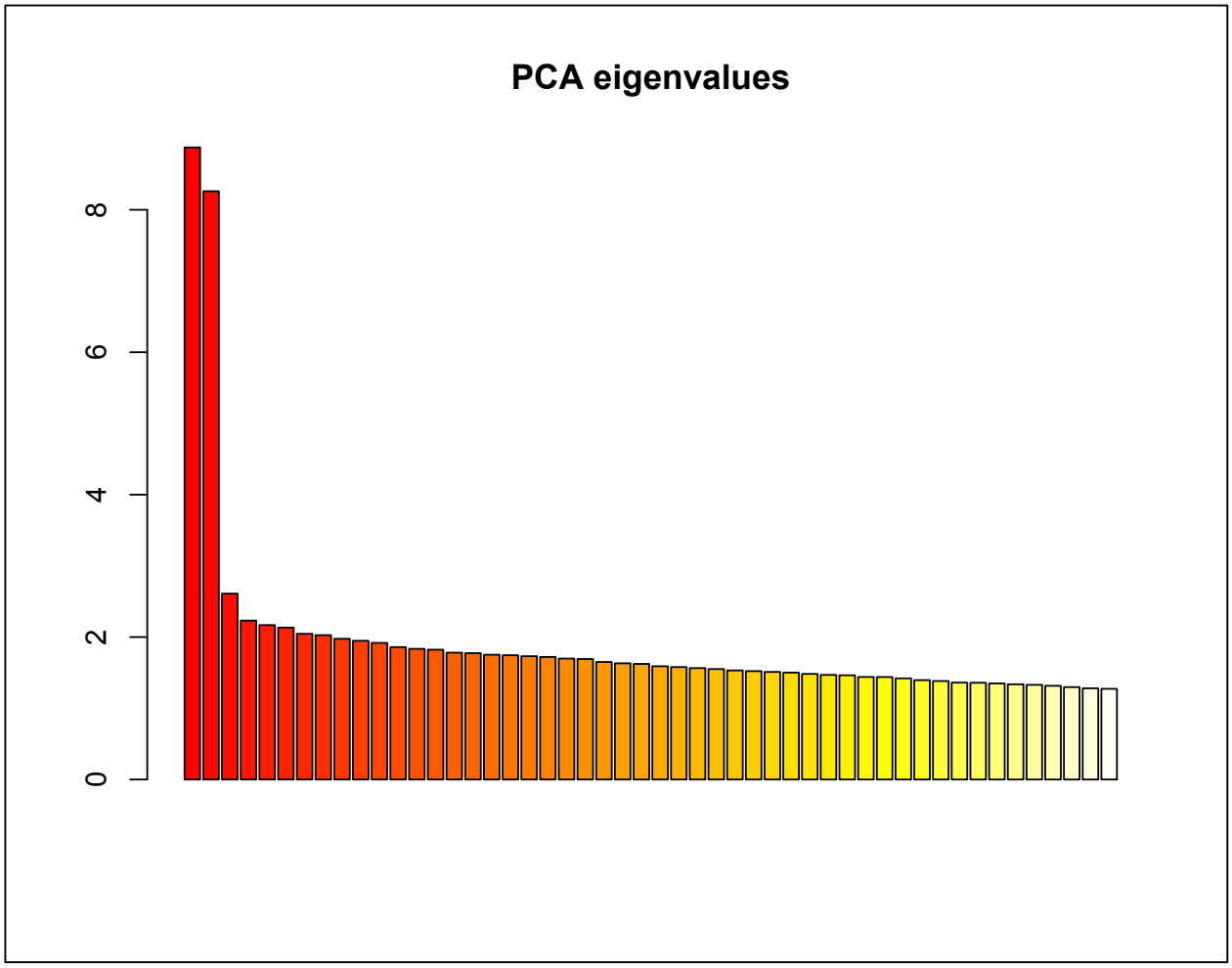

Figure 2. Plot of eigenvalues for each principal component.

.

In this PCA, all the individuals plotted extremely tight together except for individuals 0818 and 0812 who were divergent (Figure 3). These individuals obscured visualization of any groupings among the remainder of the individuals, so it was decided to repeat the PCA with individuals 0818 and 0812 excluded.

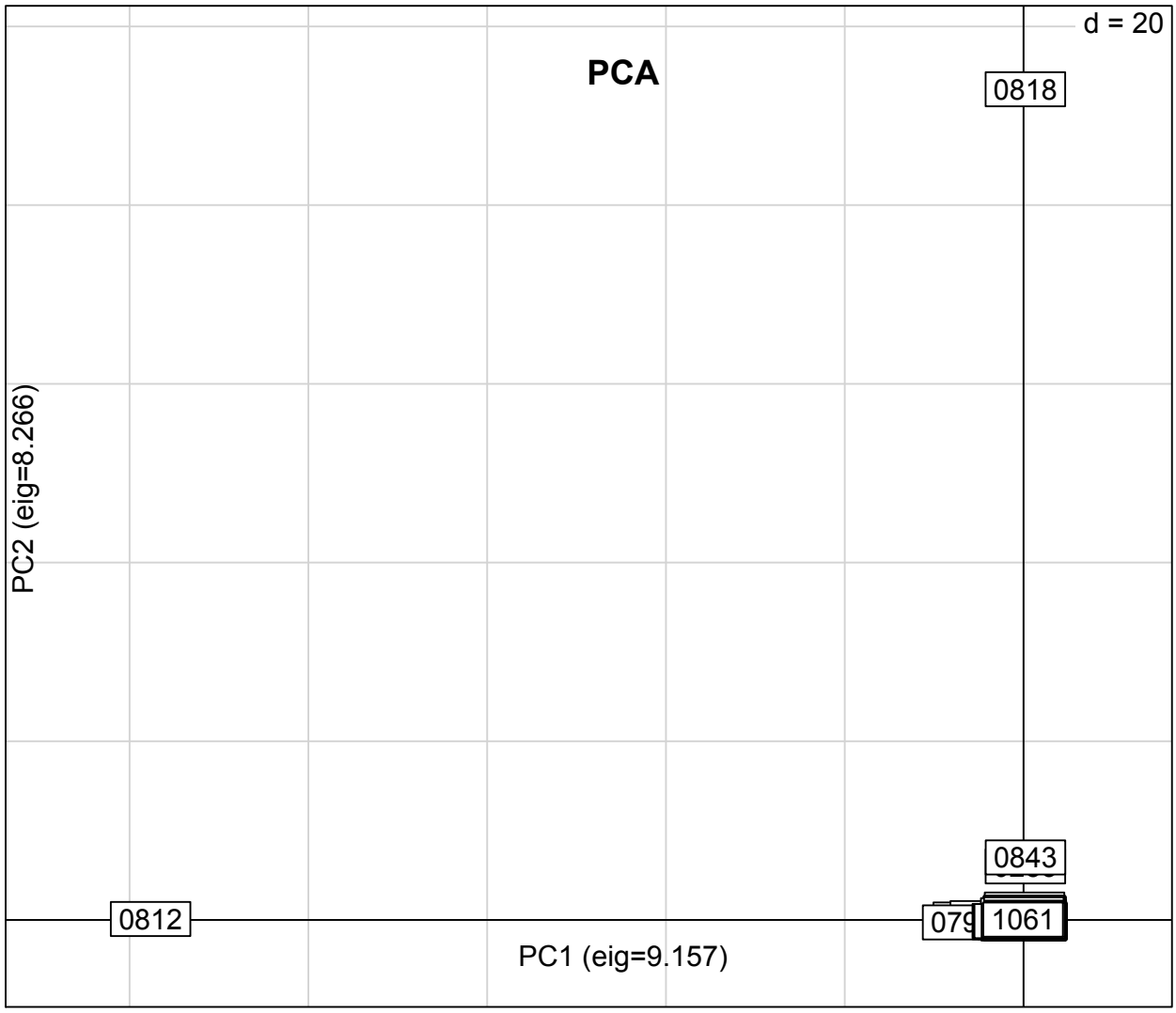

Figure 3. PCA plot showing individuals sampled.

After removing individuals 0818 and 0812, the eigenvalue for the first PC was 12.842 and the second PC's eigenvalue was 0.769. In this plot, each dot represents an individual. Circles were placed around the grouping for each city and labels were applied to the groupings' centers providing an indication of the cities' relationships to each other (Figure 4). There is no apparent separation between the cities in this plot.

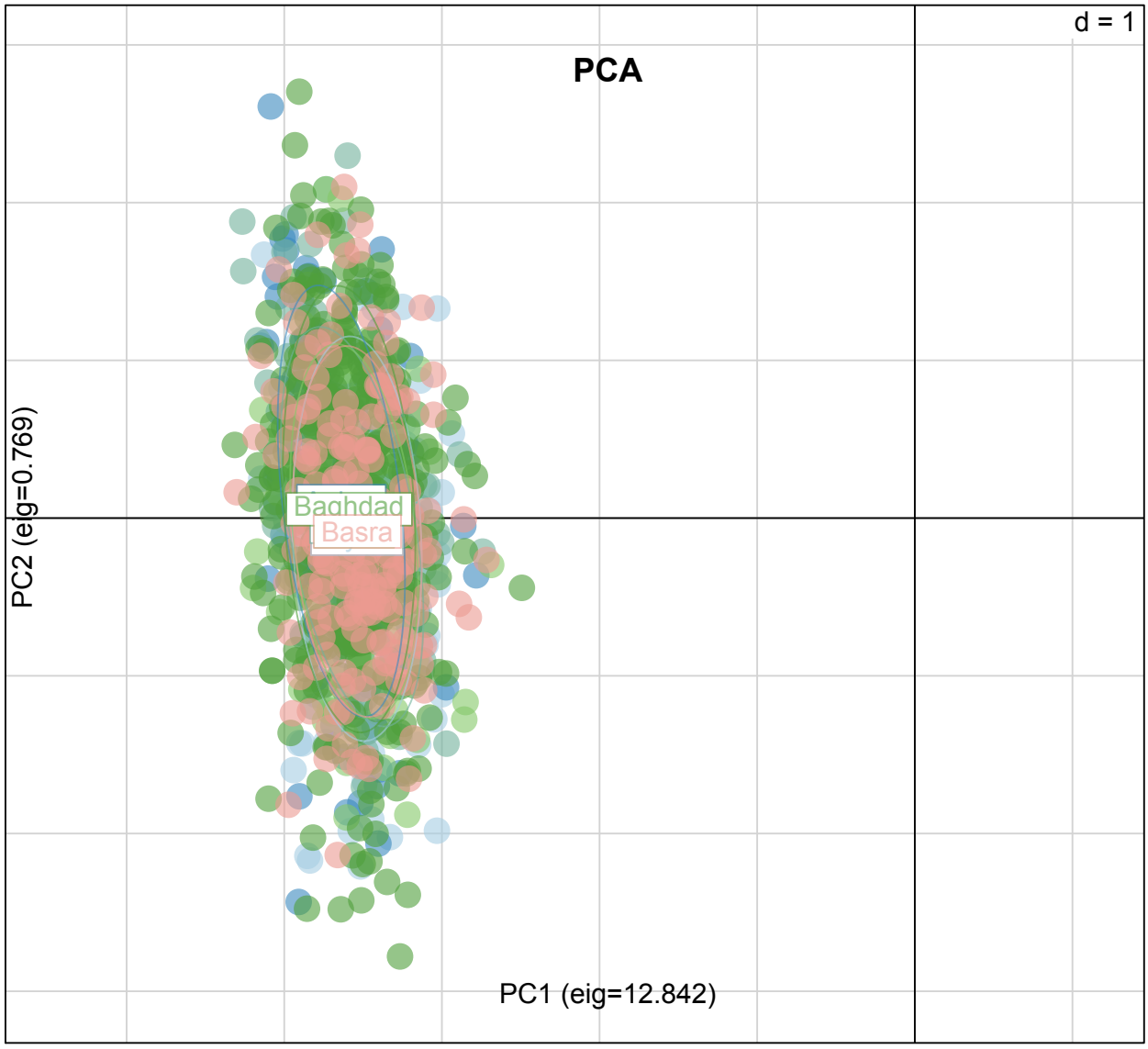

Figure 4. PCA plot with individuals 0812 and 0818 removed.

A PCA of the Iraqi cities' allele frequencies (Figure 5) revealed that Baghdad plotted near the center with Basra as the next closest city and then Diyala. Anbar plotted the furthest from the center, followed by Wasit and then Najaf. Diyala plotted closer to Wasit than to either Anbar or Najaf.

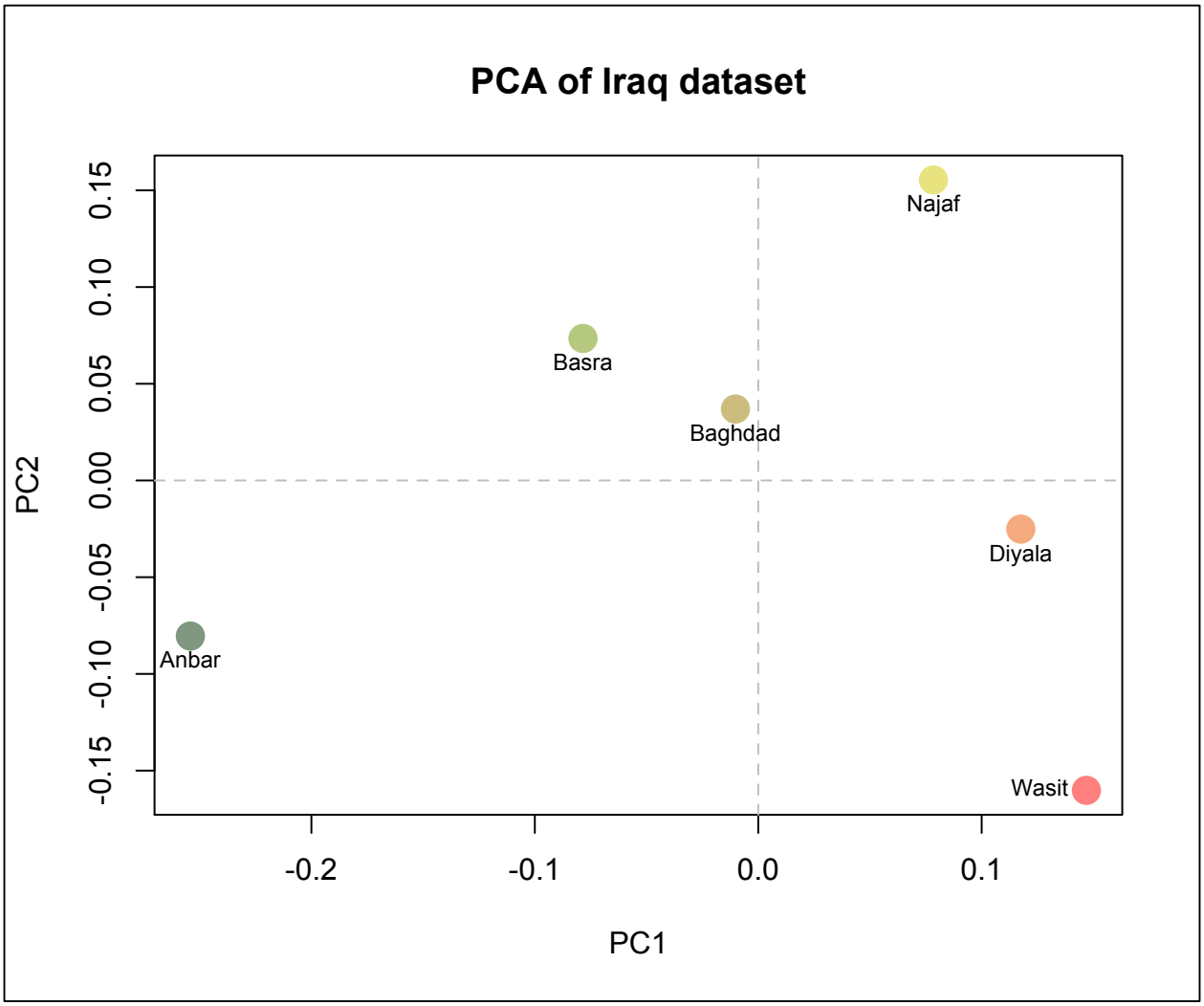

Figure 5. PCA plot of the six sampled cities in Iraq.

The PCA of Iraq's allele frequencies compared to those of surrounding countries (Figure 6) reveals that Iraq plots closest to Turkey and then Iran. Along the first principal component, Syria was plotted the furthest away Iraq with Saudi Arabia and Kuwait halfway between Iraq and Syria. The second principal component separated Saudi Arabia and Kuwait from the rest of the countries with Saudi Arabia most distant.

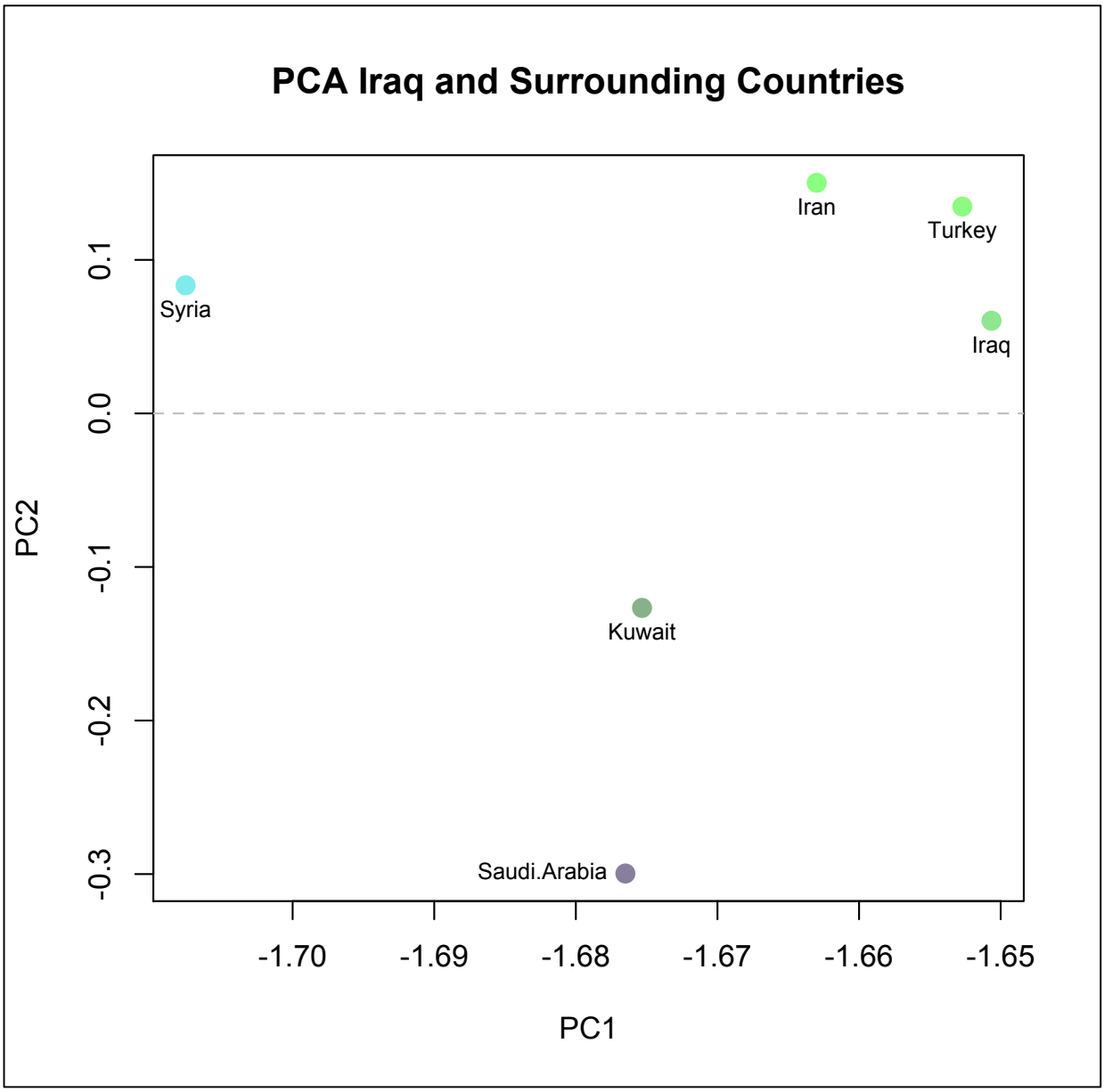

Figure 6. PCA plot of allele frequencies of Iraq and surrounding countries.

Analysis of Molecular Variance

Analysis of molecular variance (AMOVA) was performed using the "poppr" package version 2.2.0 (Kamvar, Brooks, & Grünwald, 2015; Kamvar, Tabima, & Grünwald, 2014). The results of the AMOVA found that 0.159% of the total genetic variation at these loci is between populations (cities), 2.408% of the variation is between samples, and 97.434% of the variation is within samples (Table 4).

To obtain *p*-values for these values, the *randtest* function of the "ade4" package (Chessel, Dufour, & Thioulouse, 2004; Dray & Dufour, 2007; Dray, Dufour, & Chessel, 2007) was used to perform a randomization test with 999 permutations and plotted using the *plot* function of the "graphics" package (R Core Team, 2016). The plot shows the distribution of sigma values for the 999 randomized permutations and the sigma value for the dataset is marked as a black diamond (Figure 7). The sigma value for the within-sample variation in the dataset falls outside the level of non-significance to the left of the distribution, while the between-sample and between-population sigma values fall outside the level of non-significance to the right of the distributions (Figure 7). There was significantly less variation within samples ($p = 0.001$) and significantly more variation between samples ($p = 0.001$) and between populations ($p = 0.001$) than would be expected to occur randomly (Table 4).

Source	**df**	**MS**	**Sigma**	**% Variation**	**Phi**	***p*-value**
Between populations	5	9.33812	0.009501111	0.1586367	0.001586367	0.001
Between samples	1055	6.123916	0.144191578	2.4075164	0.024113417	0.001
Within samples	1061	5.835533	5.835532516	97.4338469	0.025661531	0.001

Table 4. Results of the AMOVA test.

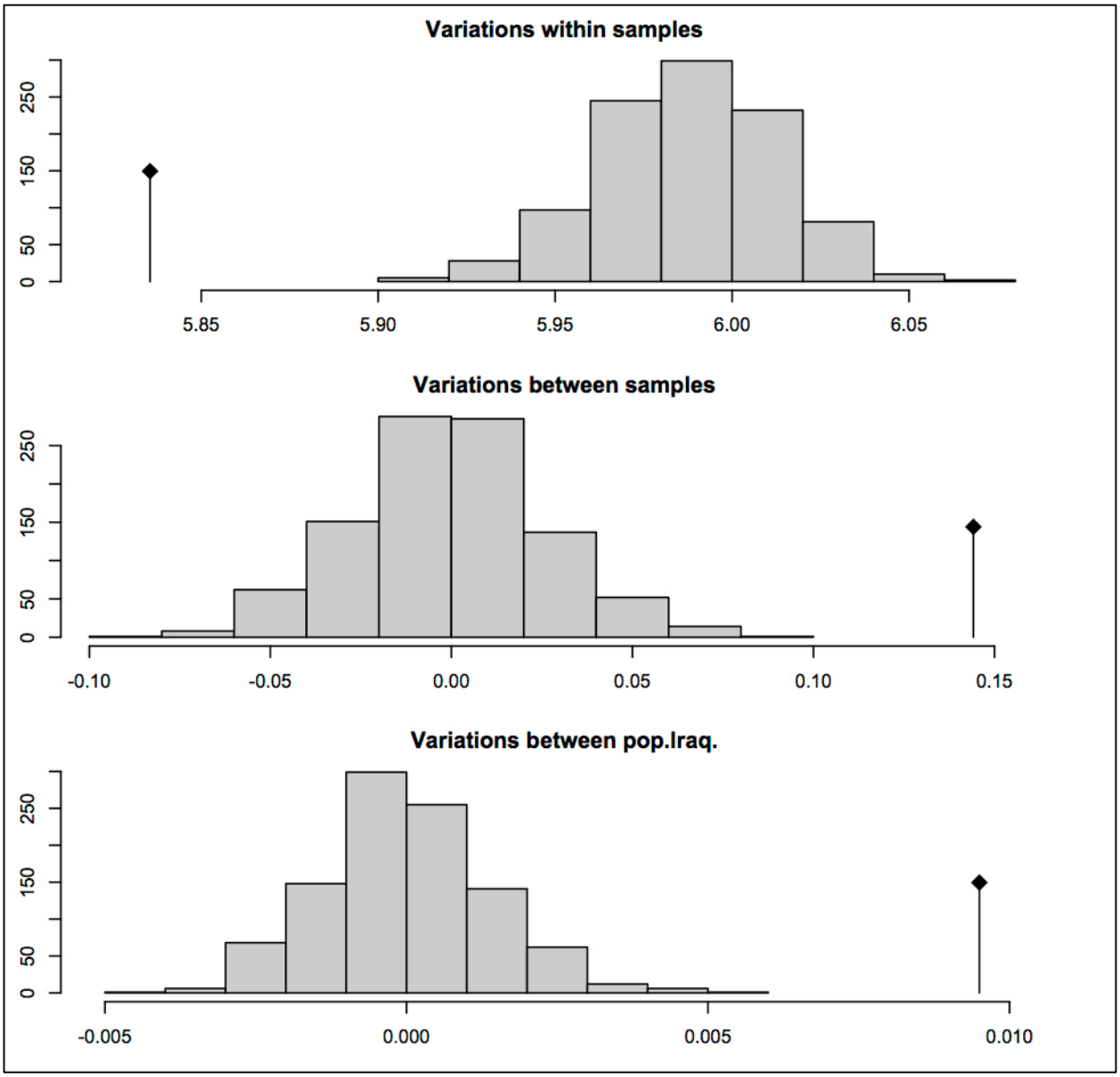

Figure 7. Results of the AMOVA randomization test.

Discriminant Analysis of Principal Components

Discriminant analysis of principal components was performed using the "adegenet" package (Jombart, 2008; Jombart & Ahmed, 2011). The function *find.clusters* was used to infer genetic clusters within the data. Based on the output, it was decided to retain 200 PCs to include all of them (Figure 8) and specified 8 genetics clusters, which had the lowest BIC score (Table 5 and Figure 9).

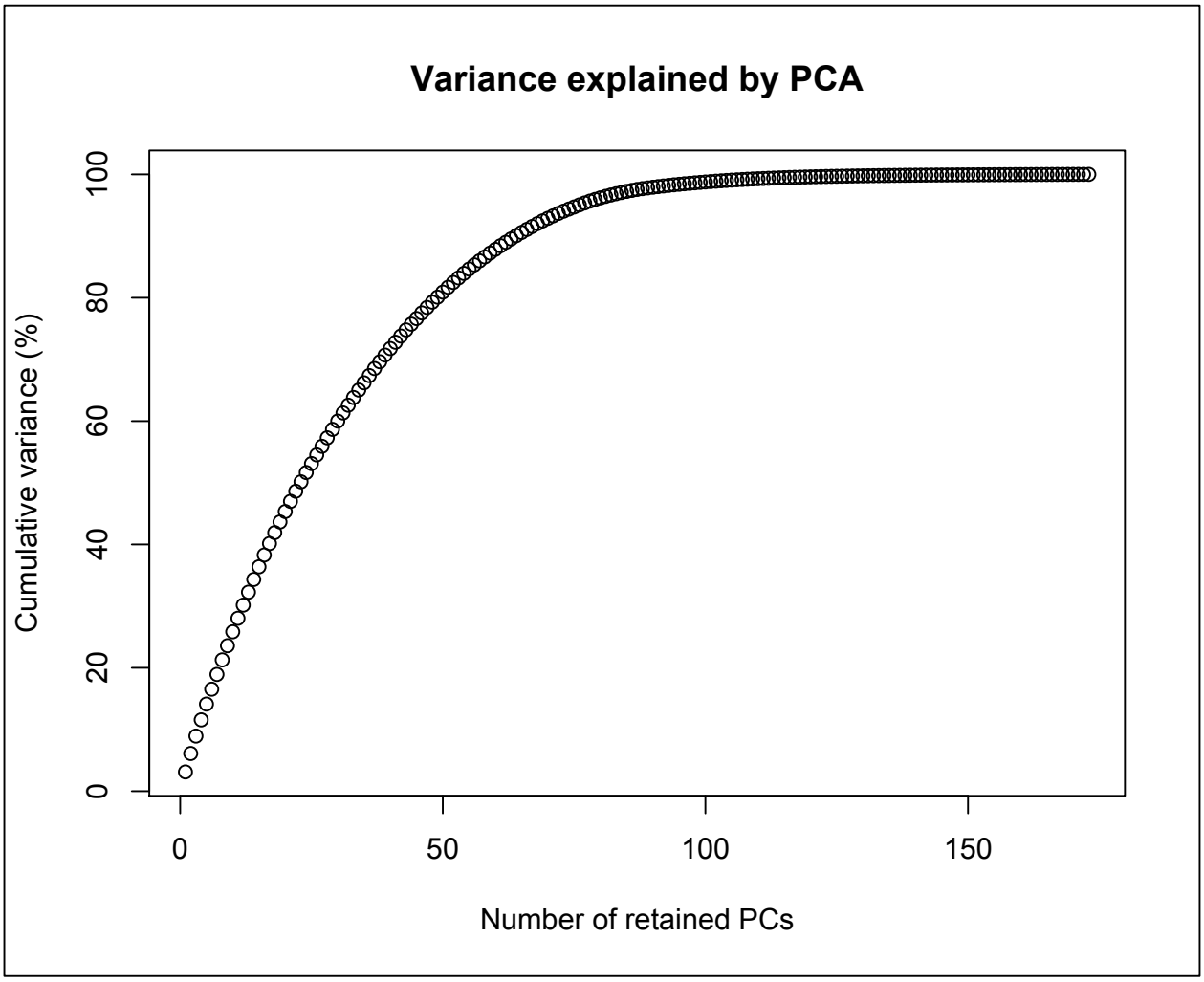

Figure 8. The cumulative percent of variance explained per principal component retained.

#	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9
BIC	1931.3	1915.6	1905.2	1899.0	1895.4	1892.4	1891.6	1890.3	1890.5

#	K=10	K=11	K=12	K=13	K=14	K=15	K=16	K=17	K=18
BIC	1891.6	1891.8	1893.4	1894.6	1895.6	1897.7	1899.1	1902.4	1904.1

#	K=19	K=20	K=21	K=22	K=23	K=24	K=25	K=26	K=27
BIC	1905.9	1908.7	1911.9	1915.8	1918.9	1921.6	1923.9	1927.2	1929.8

#	K=28	K=29	K=30	K=31	K=32	K=33	K=34	K=35	K=36
BIC	1934.1	1936.9	1942.1	1943.9	1945.4	1950.9	1955.2	1958.8	1962.2

#	K=37	K=38	K=39	K=40
BIC	1966.8	1970.9	1972.8	1978.4

Table 5. The Bayesian information criteria (BIC) for each number of clusters.

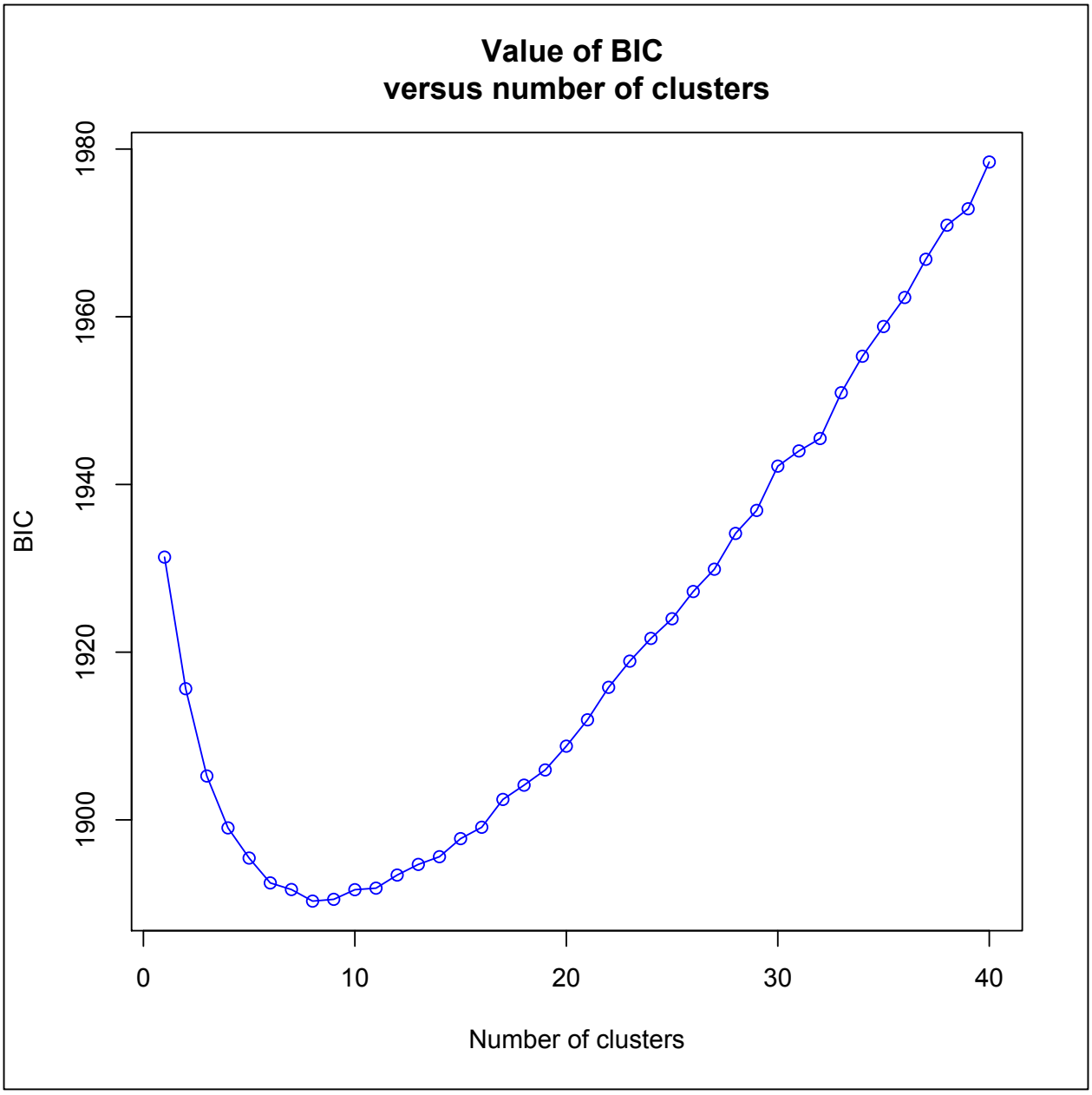

Figure 9. The Bayesian information criteria (BIC) for each number of clusters.

The output included a table of the number of individuals from each city that were assigned to each genetic cluster (Table 6). A χ^2 test of independence was performed to test whether an individual's city of residence was independent of the genetic cluster they were assigned to. The results of that test indicate that they are not independent of each other (p = 0.008).

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**	**Cluster 8**
Diyala	22	16	20	21	17	15	14	14
Anbar	16	13	9	13	15	21	32	13
Wasit	19	11	23	14	8	17	17	11
Najaf	19	13	18	13	21	13	10	11
Baghdad	48	56	46	44	39	29	53	39
Basra	17	19	22	28	31	36	25	20

Table 6. The number of individuals from each city that were assigned to each cluster.

The *dapc* function of "adegenet" was then used on the *find.clusters* output to perform a discriminant analysis of principal components. 75 PCs and 7 discriminant functions (DFs) were retained. 75 PCs were selected based on the output of the variance explained by number of PCs retained which indicated that most variance was within the first 75 PCs (Figure 8). All DFs were retained. The *dapc* output was then plotted through the "ade4" *scatter* function (Chessel, Dufour, & Thioulouse, 2004; Dray & Dufour, 2007; Dray, Dufour, & Chessel, 2007). This plot shows that although there is overlap among the individuals in the genetic clusters, the centers show no overlap and are well enough dispersed to be considered distinct (Figure 10).

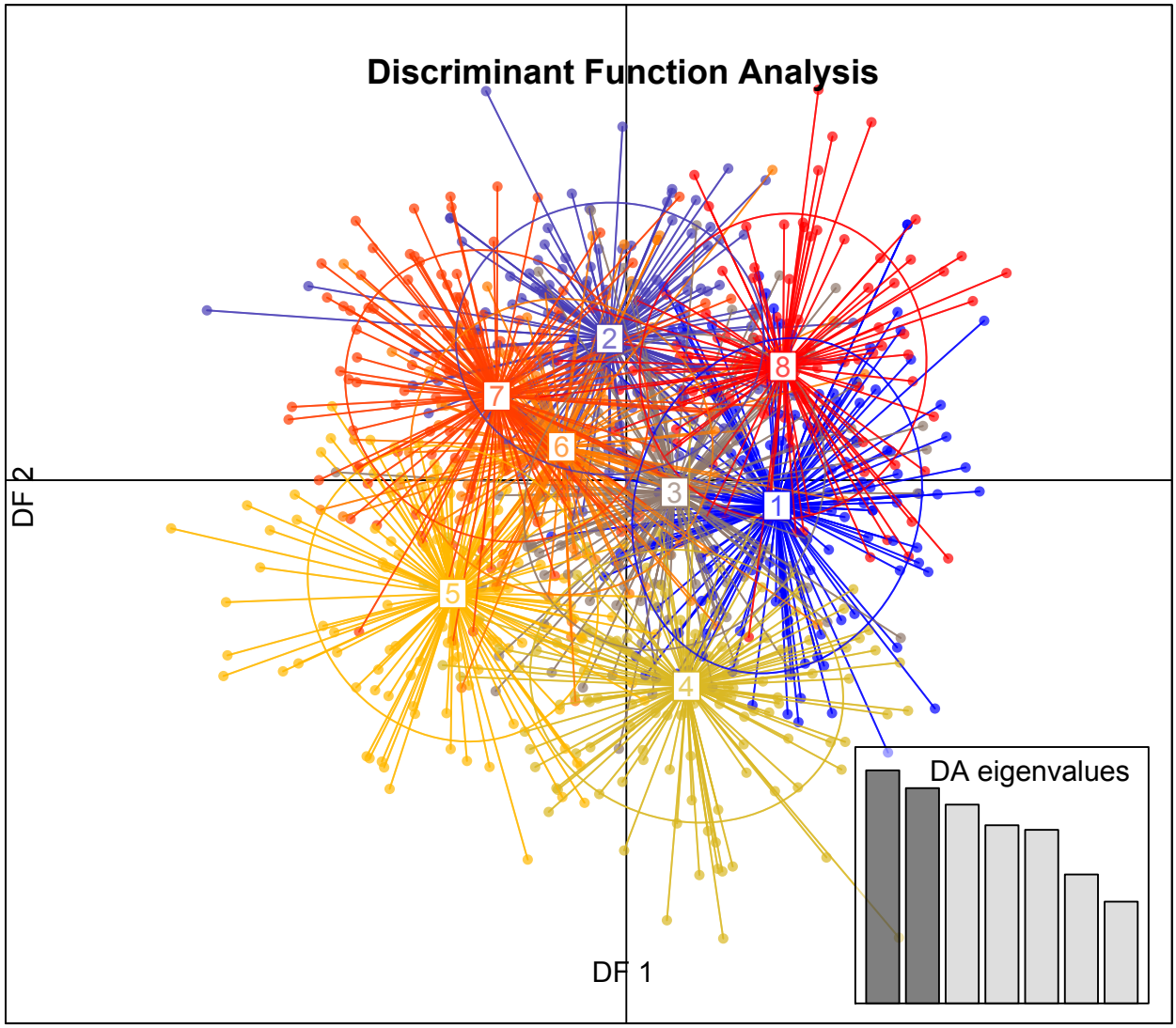

Figure 10. DAPC plot of inferred genetic clusters.

The data were evaluated to determine the optimal number of principal components (PCs) to be retained for the discriminant analysis. This was done by two methods: a-score optimization and DAPC cross-validation. The *dapc* function was performed on the raw data (with individuals grouped by cities rather than genetic clusters) retaining 75 PCs and all (5) discriminant functions. For the a-score optimization, the *optim.a.score* function from "adegenet" was used on the *dapc* output starting with 50 PCs and repeated at increments of 50 PCs until an output was obtained that did not have the maximum number of PCs retained as the optimal number. This was at 200 PCs retained which returned an optimal number of PCs as 120 (Figure 11).

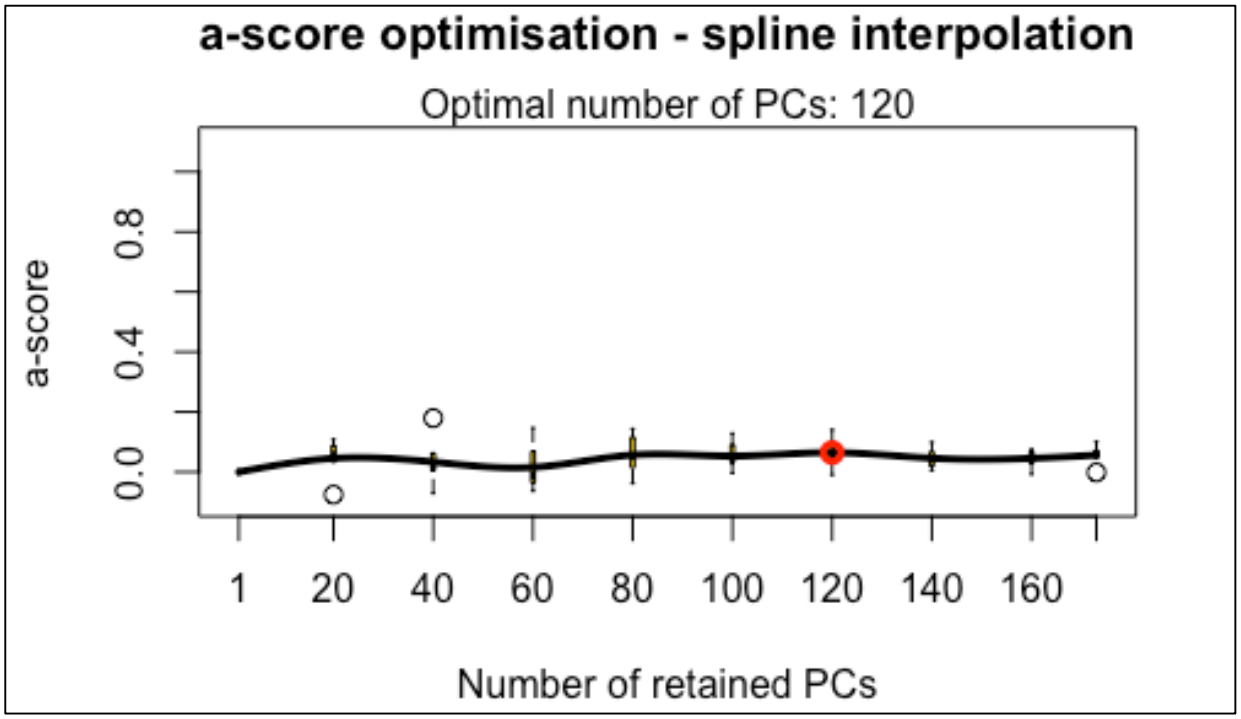

Figure 11. Plot of the calculated a-score per every 20 retained PCs.

For the cross-validation, the *xvalDapc* function from "adegenet" was used on the dataset after converting it to a matrix. It performed DAPC cross-validation starting with 20 PCs and repeated at increments of 20 PCs through 160 PCs. The output provided a plot of the proportion of successful outcomes predicted per number of PCs retained (Figure 12), a table of the mean successful assignment per number of PCs retained (Table 7) and the root mean squared error per number of PCs retained (Table 8). The best number of PCs to retain would be the one with the highest proportion of successful outcomes predicted, the highest mean successful assignment, and the lowest root mean squared error. This was 120 PCs for all three. As both a-score optimization and DAPC cross-validation gave 120 PCs as the optimal number of PCs to retain, the DAPC analysis was performed on the data with 120 PCs retained.

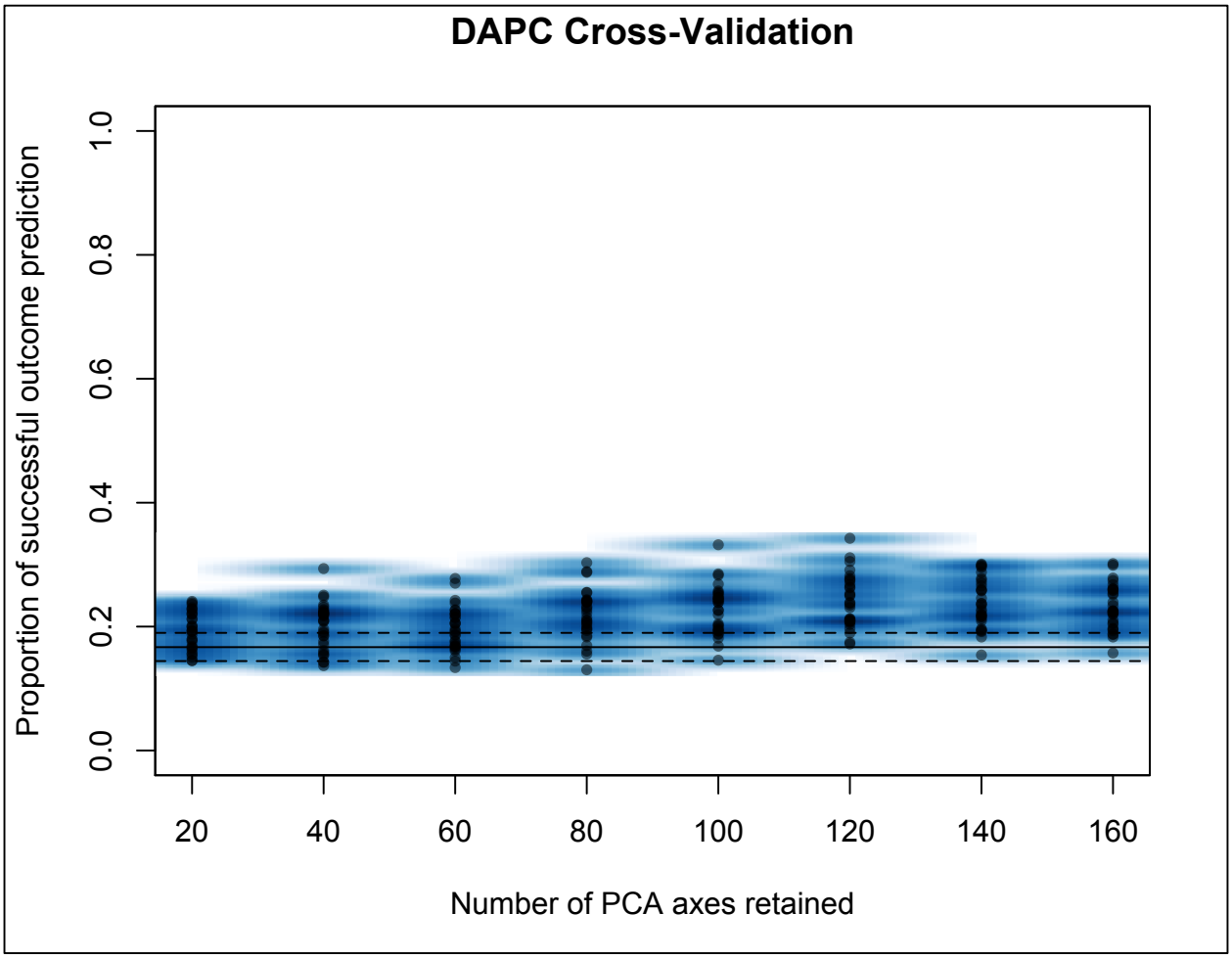

Figure 12. Proportion of successful outcomes predicted per every 20 PCs retained.

# of PCs	20	40	60	80	100	120	140	160
Mean Successful Assignment	0.18882	0.19131	0.20773	0.21424	0.22855	0.25373	0.25096	0.23304

Table 7. The mean percent successful assignment per every 20 PCs retained.

# of PCs	20	40	60	80	100	120	140	160
Root Mean Squared Error	0.81179	0.80918	0.79323	0.78648	0.77244	0.74734	0.74998	0.76844

Table 8. The root mean squared error per every 20 PCs retained.

A DAPC was performed on the dataset, with cities as the *a priori* groups, retaining 120 PCs and all DFs. The *scatter* function from "ade4" (Chessel, Dufour, & Thioulouse, 2004; Dray & Dufour, 2007; Dray, Dufour, & Chessel, 2007) was used to plot the *dapc* output (Figure 13). Baghdad plotted in the middle with Basra right next to it. Wasit was somewhat to the outside, but closer than Anbar, Diyala, and Najaf which were on the outer edges of the plot.

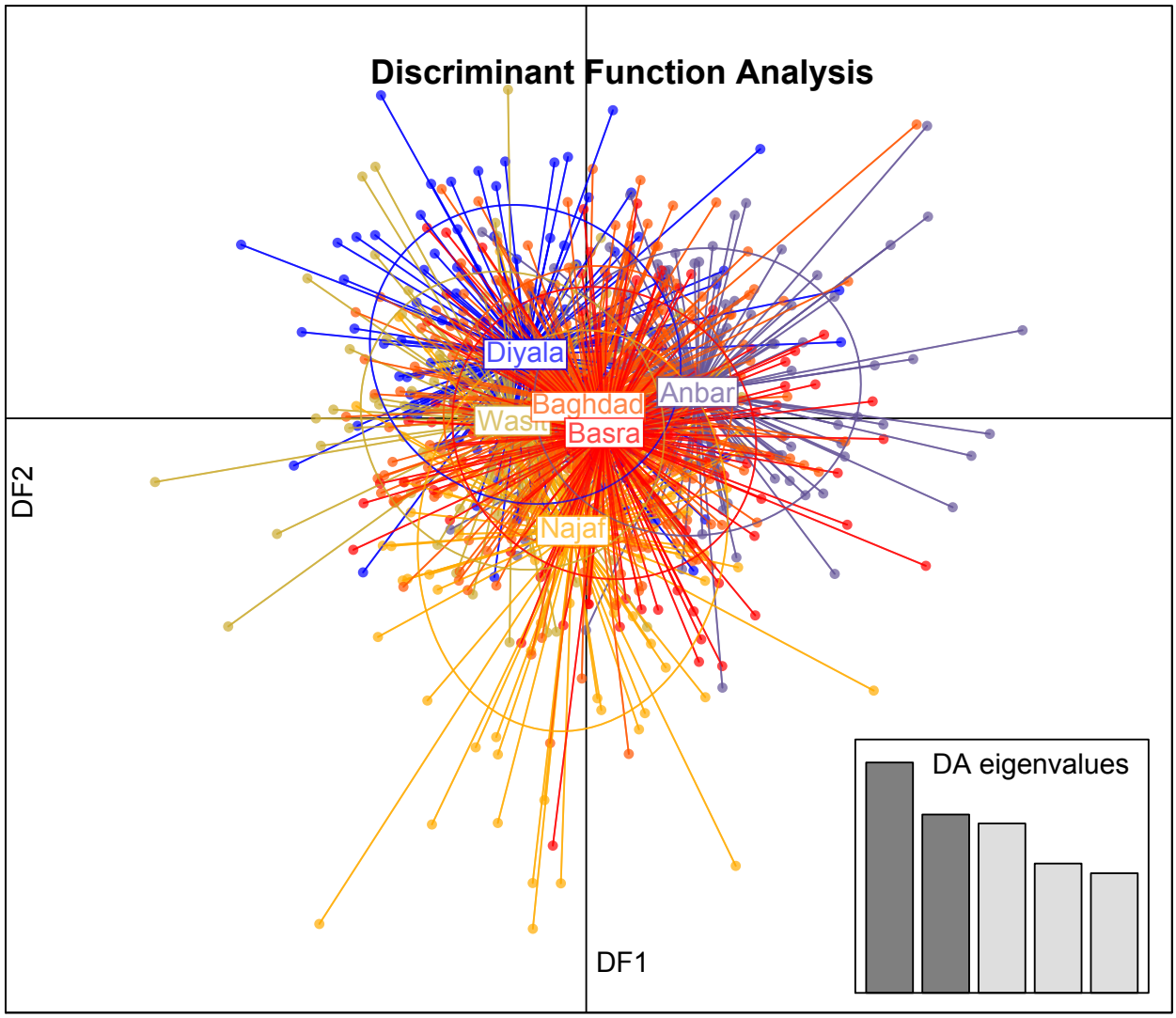

Figure 13. DAPC plot of cities.

It was found that Baghdad had the highest correct assignment rate with 69.21%. Wasit had the lowest correct assignment rate with 32.50%. Overall, 48.73% were assigned to the correct city (Table 9 and Figure 14).

City	Diyala	Anbar	Wasit	Najaf	Baghdad	Basra
% Correct	33.09353	43.18182	32.50000	36.44068	69.20904	43.93939

Table 9. DAPC percent correct assignment by city.

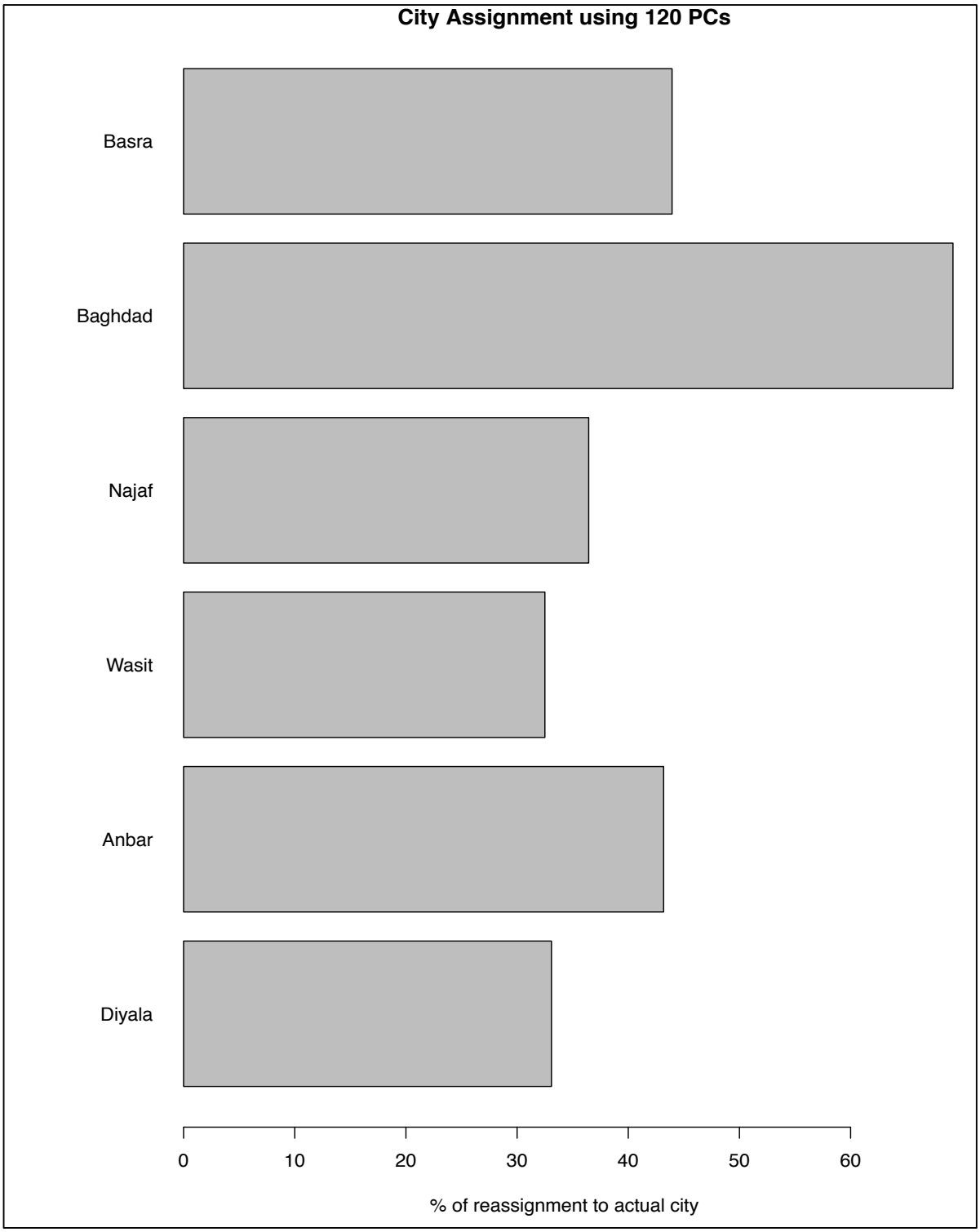

Figure 14. DAPC percent correct assignment by city.

To see how the number of PCs retained affected the percent of correct assignment, *dapc* was repeated several times at increasing increments of 5 PCs and plotted in Excel (Figure 15). Using 5 PCs, Baghdad had a correct assignment rate of 94.92% while Diyala, Wasit, and Najaf had correct assignment rates of 0.00% and Anbar and Basra had rates of 1.52% and 7.07%, respectively. As the number of PCs increased, the correct assignment to Baghdad decreased while correct assignment to all other cities increased and overall correct assignment increased.

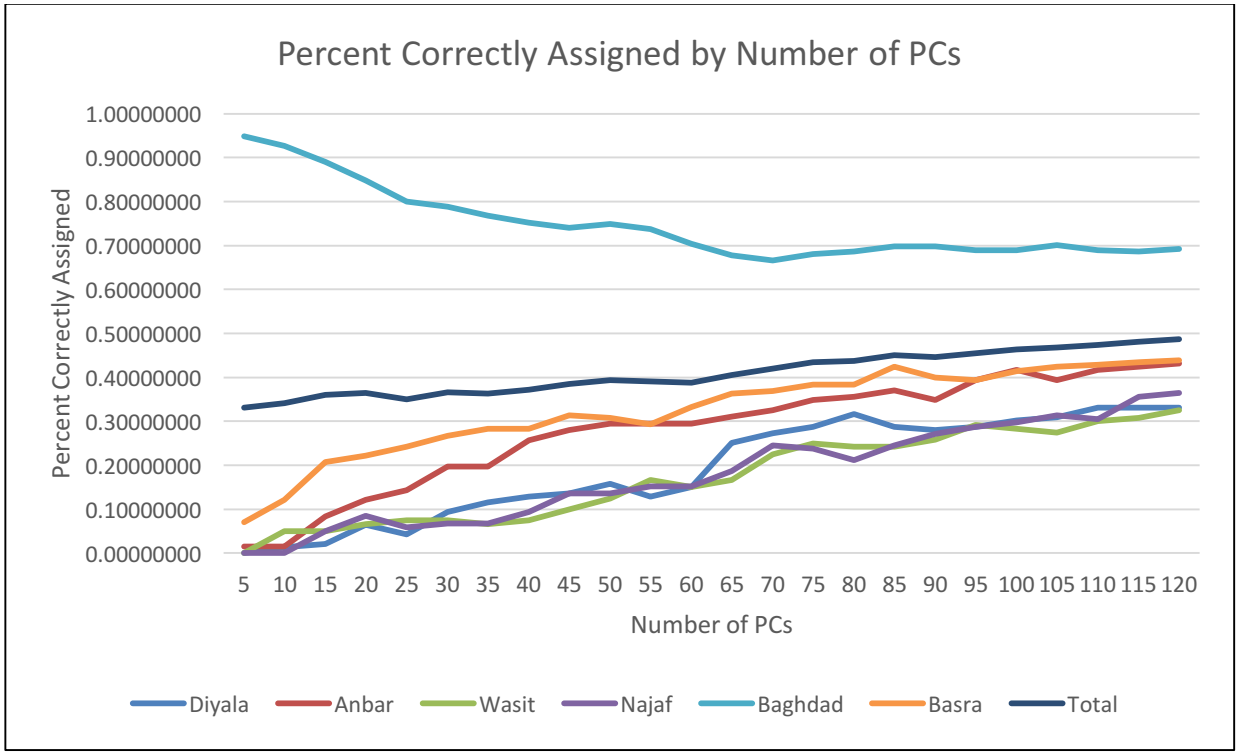

Figure 15. DAPC percent correct assignment to cities per every 5 PCs.

Multidimensional Scaling

The *cmdscale* function from "stats" package was used to perform classical (metric) multidimensional scaling and *plot* from "graphics" (R Core Team, 2016) was used to visualize the MDS output. In this plot, there are four separate grouping which include countries from Europe, Asia, Africa, and the Middle East. Iraq grouped with other countries from the Middle East (particularly Turkey and Iran) which all grouped closer to countries from Europe than either Asia or Africa (Figure 16). Goodness-of-fit using the first two coordinates was 0.690639.

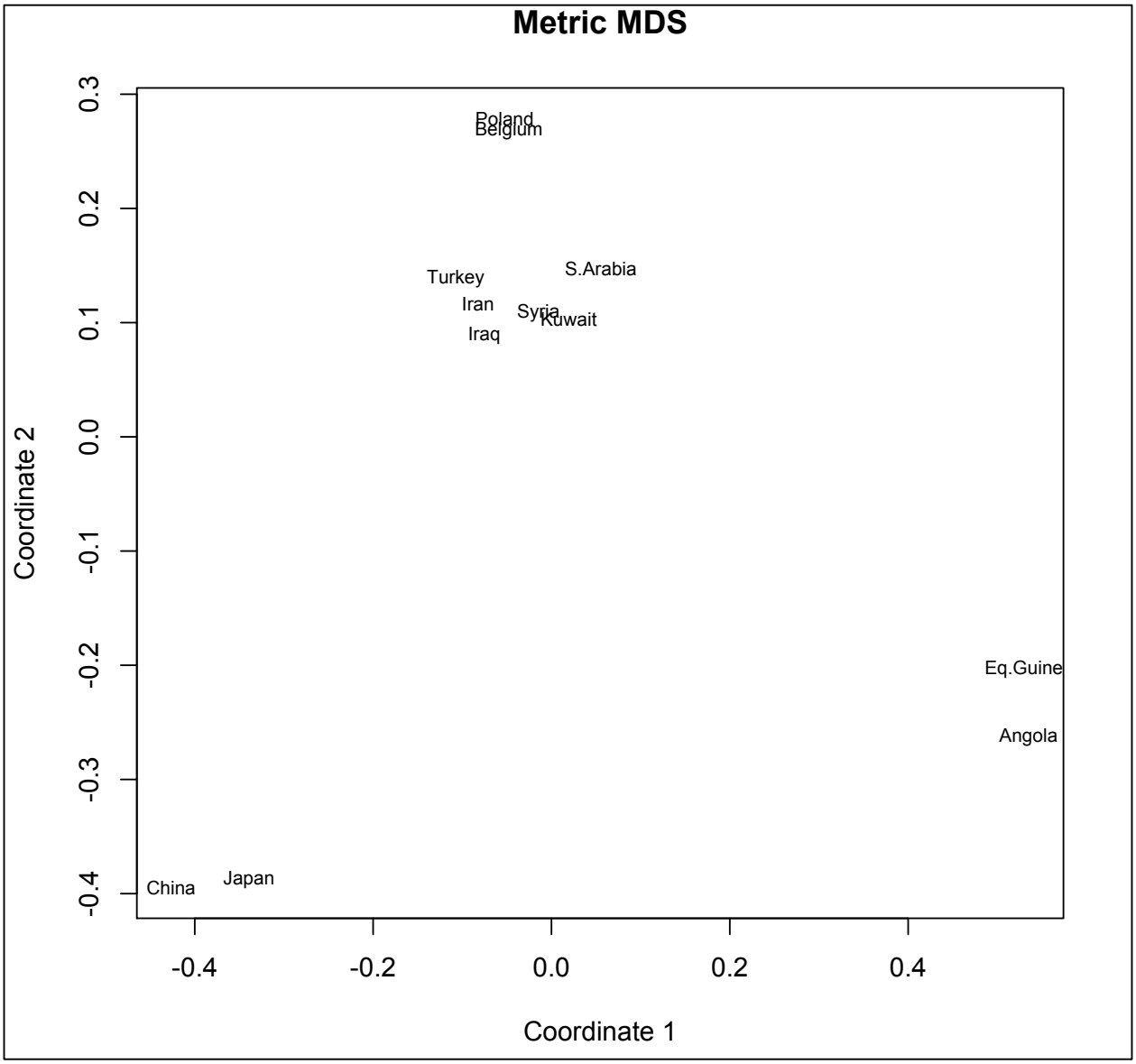

Figure 16. 2D MDS plot of countries from the Middle East, Europe, Asia, and Africa.

Adding in the third coordinate increases the goodness-of-fit to 0.798113. The z-axis further separates the Middle Eastern countries with Turkey, Iran, and Iraq forming a group on one end and Saudi Arabia by itself on the other end (Figure 17). Syria and Kuwait form a group in between. Turkey, Iran, and Iraq are also plotted closer to the European countries on the z-axis while Saudi Arabia is plotted furthest. The legend on the right of the plot provides color-coding according to the z-axis values. Points that are close to zero on the z-axis are green, very positive values are red, very negative values are blue.

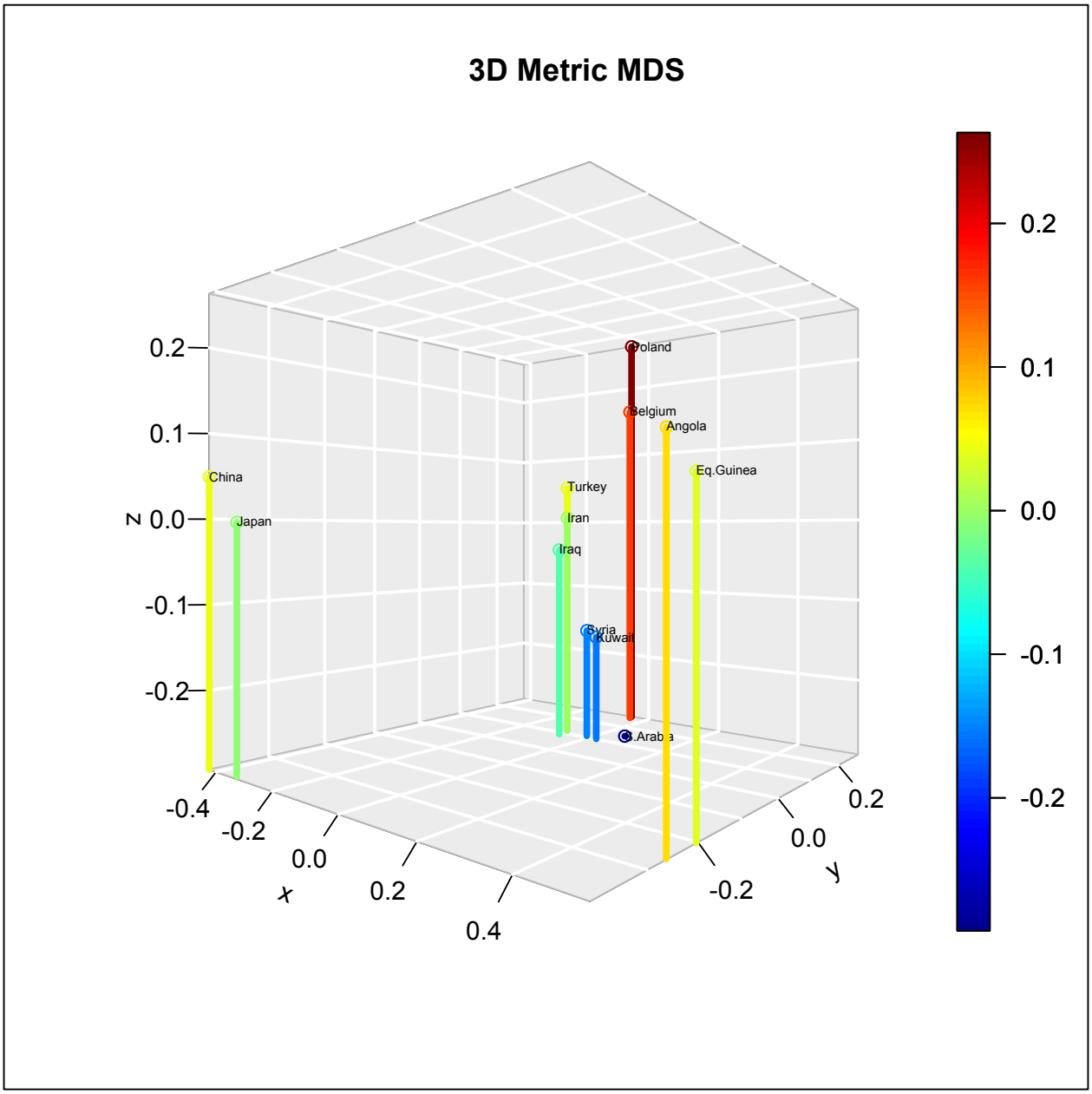

Figure 17. 3D MDS plot of countries from the Middle East, Europe, Asia, and Africa.

Forensic Applications

Allele frequencies, heterozygosities, homozygosities, matching probabilities, power of discrimination (PD), polymorphism information content (PIC), power of exclusion (PE), and the typical paternity index (TPI) at each locus were calculated for each city using PowerStats v1.2 (Tereba, 1999). This information for the whole population is included in Tables 10 and 11. The overall homozygosity for this population across all loci is 0.221 and heterozygosity is 0.779.

For forensic purposes, the most informative measures are the PIC, PD, and MP values. The PIC was above 0.60 for all loci, with TPOX having the lowest at 0.61 and D18S51 having the highest at 0.86. The average PIC was 0.77. The PD was above 0.8 for all loci, with TPOX having the lowest at 0.835 and D2S1338 having the highest at 0.974. The average PD was 0.927. The matching probability, the probability of obtaining a match among randomly selected individuals, was lowest for D2S1338 (0.026) and highest for TPOX (0.165), with an average MP of 0.073. Multiplying together the MPs for each locus resulted in a combined MP (CMP) of 1.227×10^{-18}. Another way of expressing this is that matching genotypes using these loci will be found, on average, in 1 person in 8.152×10^{17}.

For paternity cases, the important measures are the PE and TPI. The power of exclusion was lowest for TPOX (0.384) and highest for FGA (0.682). The average PE was 0.563. The total power of exclusion (TPE), or the probability of excluding an unrelated individual using all 15 loci, was 0.999997. The locus with the smallest TPI was TPOX (1.52) and the locus with the largest TPI was FGA (3.19). The average TPI was 2.36. The combined paternity index (CPI) for all 15 loci was 289,622. A matching profile at all loci is 289,622 times more likely to belong to the biological father of the child than to a randomly selected individual.

	D8S1179	**D21S11**	**D7S820**	**CSF1PO**	**D3S1358**	**TH01**	**D13S317**	**D16S539**
5				0.000		0.000		
6						0.280		0.000
7			0.016	0.002	0.000	0.184		0.000
8	0.009		0.165	0.004		0.127	0.149	0.040
8.3						0.000		
9	0.006		0.109	0.026		0.248	0.071	0.171
9.3						0.136		
10	0.077		0.267	0.269		0.022	0.069	0.084
11	0.081		0.262	0.307	0.000	0.000	0.294	0.306
12	0.104		0.159	0.329	0.001		0.304	0.241
13	0.267		0.020	0.052	0.003		0.076	0.136

	D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539
14	0.193		0.002	0.008	0.055		0.033	0.020
15	0.202			0.001	0.267		0.002	0.000
16	0.053			0.000	0.275	0.000	0.001	
17	0.007				0.253		0.000	
18	0.002				0.135			
19					0.008			0.000
20					0.001			
22							0.000	
25								0.000
26		0.003						0.000
27		0.012						
28		0.141						
29		0.230						
29.2		0.005						
30		0.235						
30.2		0.025						
31		0.045						
31.2		0.108						
32		0.005						
32.2		0.135						
33		0.002						
33.2		0.047						
34		0.000						
34.2		0.005						
35		0.001						
36		0.000						
MP	0.052	0.048	0.072	0.130	0.091	0.076	0.079	0.075
MP expressed as 1 in …	19.1	20.8	13.9	7.7	10.9	13.1	12.7	13.3
PD	0.948	0.952	0.928	0.870	0.909	0.924	0.921	0.925
PIC	0.80	0.82	0.76	0.67	0.73	0.76	0.75	0.76
PE	0.619	0.647	0.561	0.439	0.513	0.531	0.570	0.543
TPI	2.64	2.87	2.27	1.71	2.02	2.11	2.32	2.17
Ho	0.189	0.174	0.221	0.293	0.248	0.238	0.216	0.231
He	0.811	0.826	0.779	0.707	0.752	0.762	0.784	0.769

Table 10. Forensic measures for 8 of the 15 STRs.

	D2S1338	D19S433	vWA	TPOX	D18S51	D5S818	FGA
6				0.003			
7				0.002	0.000		
8	0.000		0.001	0.510		0.008	
9				0.127	0.002	0.057	
9.2		0.000					
10	0.000	0.001	0.000	0.091	0.008	0.102	
10.2					0.000		
11		0.008	0.001	0.236	0.019	0.308	
12	0.000	0.089	0.000	0.029	0.132	0.319	
12.2		0.001					
13	0.001	0.234	0.004		0.174	0.194	
13.2		0.024					
14	0.001	0.251	0.077		0.173	0.011	
14.2		0.052		0.000			
15	0.000	0.139	0.107	0.000	0.135	0.001	0.000
15.2		0.094					
16	0.048	0.046	0.247		0.116		0.001
16.2		0.042			0.001		
17	0.193	0.011	0.286		0.097		
17.2		0.006					0.000
18	0.117	0.001	0.186	0.000	0.080		0.009
18.2		0.001			0.000		0.001
19	0.137		0.079		0.037		0.063
19.2							0.001
20	0.141		0.012	0.000	0.012		0.095
20.2							0.000
21	0.049				0.008		0.167
21.2							0.005
22	0.041				0.002		0.148
22.2							0.004
23	0.124				0.000		0.155
23.2							0.002
24	0.077				0.001		0.183
24.2							0.002
25	0.057		0.000				0.101
26	0.010						0.043
27	0.000						0.006
28	0.000						0.004

	D2S1338	D19S433	vWA	TPOX	D18S51	D5S818	FGA
29							0.006
32							0.000
MP	0.026	0.045	0.068	0.165	0.029	0.102	0.033
MP expressed as 1 in …	37.8	22.3	14.7	6.0	34.5	9.8	30.4
PD	0.974	0.955	0.932	0.835	0.971	0.898	0.967
PIC	0.87	0.82	0.77	0.61	0.86	0.71	0.85
PE	0.669	0.629	0.520	0.384	0.646	0.497	0.682
TPI	3.07	2.72	2.05	1.52	2.85	1.94	3.19
Ho	0.163	0.184	0.244	0.330	0.175	0.257	0.157
He	0.837	0.816	0.756	0.670	0.825	0.743	0.843

Table 11. Forensic measures for remaining 7 STRs.

Chapter 5: Discussion

Genetic Structure of Iraq

Hardy-Weinberg Equilibrium and F-Statistics

The Hardy-Weinberg principle states that the gene frequencies within a population will remain constant across generation in the absence of evolutionary forces such as natural selection, genetic drift, gene flow, and mutation (Castle, 1903; Hardy, 1908; Weinberg, 1908). However, a population may be in Hardy-Weinberg equilibrium (HWE) without all the assumptions being valid due to forces counterbalancing each other. Additionally, violations from HWE may result from sampling error, population substructure, misclassification of genotypes, or failure to detect rare alleles.

Wright's F-statistics measures the extent of differentiation among subpopulations by estimating the average deviation of the genotypic proportions from HWE (Wright, 1951). There are three *F*-statistics that examine different levels of population structure: an individual (*I*) to the total (*T*) population, F_{IT}, and individual (*I*) to the subpopulation (*S*), F_{IS}, and a subpopulation (*S*) to the total (*T*) population, F_{ST}. These statistics provide information regarding inbreeding within the population and whether there is genetic differentiation between the subpopulations.

Together, HWE and *F*-statistics can be used to test for the Wahlund effect, in which a reduced heterozygosity is seen as a result of population stratification (Wahlund, 1928). Specifically, if two or more of the subpopulations differ in their allele frequencies ($F_{ST} >> 0$), the total heterozygosity will be reduced ($F_{IT} >> 0$), even if the subpopulations themselves are in HWE ($F_{IS} = 0$) (Dharmarajan, Beatty, & Rhodes Jr, 2012).

F_{ST} measures the mean reduction in heterozygosity of a subpopulation relative to the total population. It can range from 0 to 1 with 0 indicating no differentiation and 1 indicating

complete differentiation wherein all the subpopulations are fixed for different alleles. F_{IS} measures the mean reduction in heterozygosity of an individual relative to its subpopulation. It can range from -1 to 1 with -1 being all individuals are heterozygous and 1 meaning that no individuals are heterozygous. F_{IT} measures the mean reduction in heterozygosity of an individual relative to the total population. It can also range from -1 to 1.

After correcting for multiple testing, none of the cities that were sampled deviated from HWE at any of the 15 STR loci (Table 1). The expectation then is that the F_{IS} value will be close to 0, indicating that the within-population is in HWE. This was shown to be the case, with the average F_{IS} value across all loci being 0.0231 (Table 2). If there is population stratification, then one would expect to see a large F_{ST}. F_{IT} is rarely used, but if overall heterozygosity is reduced due to a Wahlund effect, a large F_{IT} would be expected. *F*-statistics are not an absolute measure, but rather a relative measure. Originally, it was suggested that an F_{ST} of 0.00-0.05 be considered indicative of low genetic differentiation, 0.05-0.15 revealing moderate differentiation, 0.15-0.23 being great differentiation, and values above 0.23 high differentiation (Holsinger & Weir, 2009). However, it has been well established that there is low genetic diversity among humans and F_{ST} values will rarely exceed 0.05 for microsatellite data (Balloux & Lugon-Moulin, 2002).

Even taking the low genetic diversity of humans into account, the average F_{ST} value obtained for the Iraqi population is still low, only 0.0016, with an F_{IT} value of 0.02474. The low F_{ST} value, F_{IT} and F_{IS} values close to zero, and the cities all being in HWE indicate that, overall, the individuals sampled can all be considered part of a single population and that the different cities within Iraq do not represent separate populations.

Nei's Genetic Distance

The conclusion supported by the HW test and *F*-statistics, that the populations in the various cities are all genetically closely related, receives additional support from Nei's genetic distance. Nei's genetic distance is a measure of the genetic differences between two populations (Nei, 1972; Nei, 1978). It measures the accumulated number of gene differences per locus and, assuming that the rate of genetic change per year is constant, it is linearly related to the divergence time between the two populations. Populations that are identical will have a genetic distance of zero. The larger the value of *D*, the greater the genetic distance between the two populations (Nei, 1972; Nei, 1978). Microsatellite data from human populations across the world averaged a genetic distance of 0.1 (Nei & Takezaki, 1996). All the genetic distances between cities in Iraq were less than 0.03 indicating that they are all closely related (Table 3).

Even though all the genetic distances are low, it is still possible to look at which cities/regions are more closely related to each other. The greatest genetic distances are found in Anbar's relationship to other cities with the greatest distances being between Anbar and Wasit, then Anbar and Diyala, followed by Anbar and Najaf. The only genetic distances that exceed 0.02 are found in Anbar. Conversely, the lowest genetic distances are all found in Baghdad's relationships with the lowest being Baghdad and Basra, then Baghdad and Diyala, and finally Baghdad and Najaf (Table 3).

It is unlikely that Anbar's higher genetic distances is due purely to geographical separation. The city that is furthest from the other cities is Basra, which has the second lowest genetic distances. Also, Wasit has the second highest genetic distances but is proximal, geographically, to Diyala, Baghdad, and Najaf (Figure 1). It is also unlikely that the differences are due solely to ethnic differences as the majority of people living in the regions sampled

identify as Arab. We do not have ethnicity data for the individuals sampled, but it is a safe assumption that most individuals in the sample are of Arab descent. The second largest population in Iraq is Kurdish, but these individuals are mainly found in northern Iraq where samples were not taken. There are minority populations of Turkomen, Assyrian, Jewish, and others which are mostly likely to appear in Baghdad which has a more mixed population than the other cities.

What is most striking is that Anbar, which has the highest genetic distances, is situated in western Iraq which is mainly Sunni Arabs while Basra, Diyala, and Najaf are found in Eastern and Southeastern Iraq which is mainly Shia Arabs. Baghdad, with the lowest genetic distances, is in a region that has a fair mix of Sunnis and Shias (Figure 18). Wasit, which has the second highest genetic distances, is composed mainly of Shia Arabs; however, it is also located in the region of the marshlands which are home to the Marsh Arabs which have a remained a more endogenous, less admixed, population than the rest of the Arab population in Iraq (Al-Zahery, et al., 2011). It is rare for individuals to marry outside their religious denomination and, in fact, first-cousin marriage is common in Iraq, and these practices may be leading to microdifferentiation between Sunni and Shia Arabs in Iraq. However, it is important to note that 1) conversion between Sunnism and Shi'ism does occur (Elad-Altman, 2007) and 2) the majority of Iraqi Shia are from a relatively recent conversion from Sunnism that occurred during the nineteenth century (Cline, 2000). So, there is question of whether there has been enough separation and time between the two groups for differentiation to occur which can only be answered through further studies.

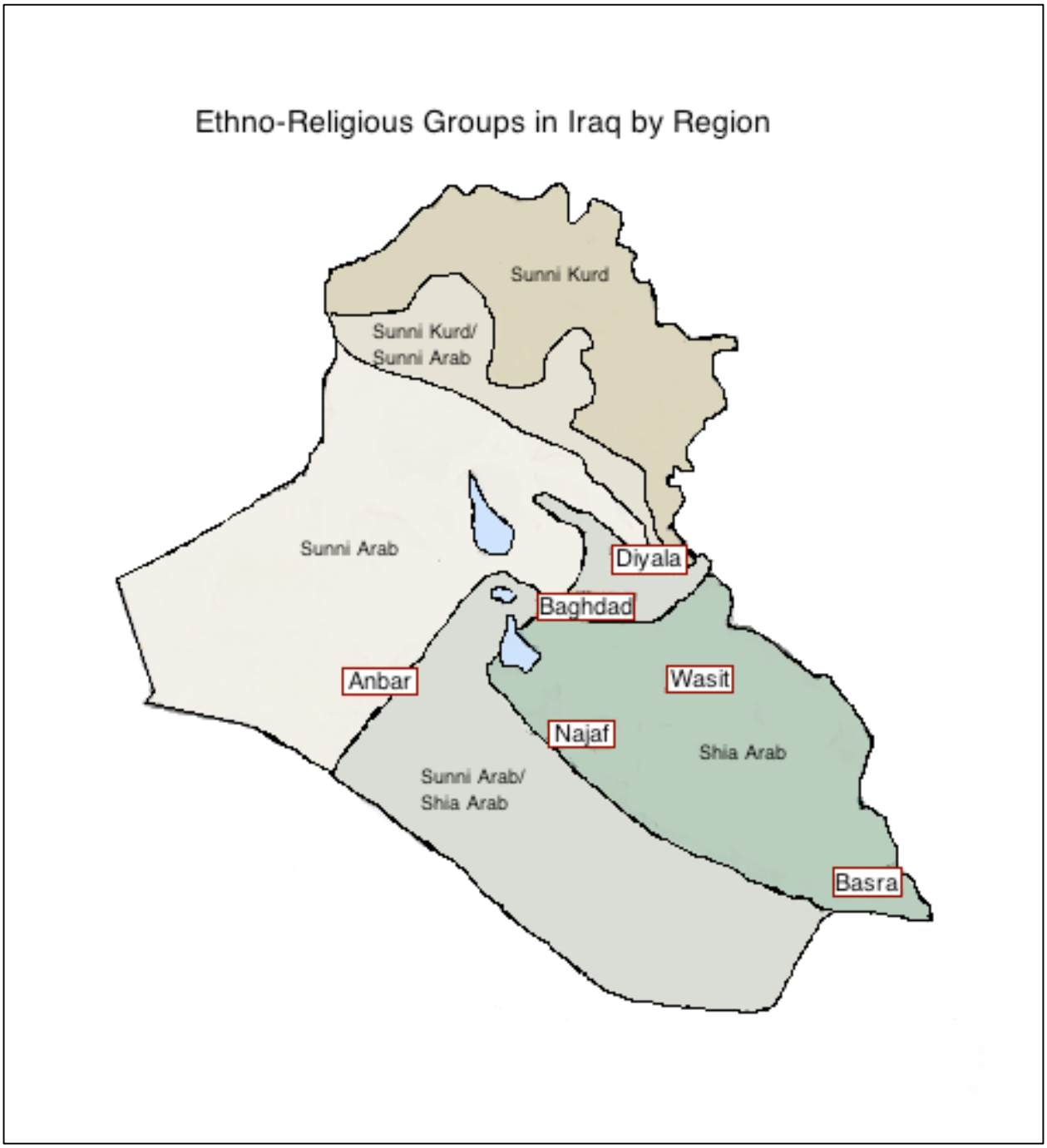

Figure 18. Ethno-Religious Distribution of Iraq.

Principal Component Analysis

Principal Component Analysis (PCA) transforms the dataset into a series of synthetic variables, called principal components (PCs), with the first PC having the largest possible variance, and each succeeding component having the highest variance possible under the restriction that each PC is orthogonal (perpendicular) to the preceding PC (Hotelling, 1933a; Hotelling, 1933b). This allows for a visual representation of the variance by allowing plotting of the data points along two or three eigenvectors. To look at population stratification among the cities in Iraq, PCA was performed on the raw dataset, in which each individual appeared as a plot point, and on the cities' allele frequencies, in which each city appeared as a plot point. A PCA was also done to compare allele frequencies at these loci between Iraq and the surrounding countries.

If the population was highly stratified, then one would expect to see distinct groupings of individuals according to the city they were sampled in. This was not the case with this dataset. The first PCA plot showed a tight grouping of all individuals except for two, who were distinct from the rest of the population and from each other (Figure 3). This could have been due to sequencing error or they may represent ethnic groups separate from the rest of the population. Regardless, they do not fit in with the rest of the population and their presence in the PCA plot obscured visualization of any relationships within the remaining individuals. Therefore, the PCA was run again with those two individuals removed.

In this PCA plot, the individuals are plotted with a circle surrounding the individuals in a city with a label designating that city centered (Figure 4). Therefore, the cities' labels provide an indication of the cities' relation to each other within the plot. Although the individuals are dispersed along the second PC, they are not as spread out along the first PC and they do not form

separate clusters according to city. There is some minor separation of the cities, as their centers are not exactly on top of each other, but there is much overlap. This supports the F-statistics and Nei's genetic distance in that it appears that the cities have only minor, if any, population stratification. However, the fact that the individuals are spread out along the second PC suggests that there is genetic diversity in Iraq, but that it is found across the cities rather forming discrete groups.

To get a better look at the relationships of the cities themselves, it is important to summarize the dataset into allele frequency data. Since there is diversity among the individuals, the centered circles were affected by the range of the plotted individuals and could therefore be impacted by individuals whose alleles may be rare and not representative of the city they are from. To make sure that the data are representative of each city, the allele frequencies were used to perform a PCA and plot the cities' relationships to each other.

The results of that PCA plot find similar distance relationships as was seen with Nei's genetic distance (Figure 5). Baghdad and Basra plot in the middle, indicating that they are closer to the more outlying cities. Anbar plotted the furthest away from the other cities with Wasit and Najaf also falling along the edges of the plot. Diyala plotted closer to Wasit. All of this is consistent with what was seen with Nei's genetic distance.

A PCA was also performed on the allele frequencies for the country of Iraq compared to surrounding countries (Figure 6). Iraq plotted closest to Iran and Turkey while Syria and Saudi Arabia plotted the furthest away from Iraq. Kuwait was positioned in the middle. This suggests that there are genetic differences between countries in the Levant and those on the Arabian Peninsula, which is consistent with what has been found in mtDNA and Y-chromosome studies on the region (Badro, et al., 2013; El-Sibai, et al., 2009). The close relationship of Iraq with both

Turkey and Iran has also been established in previous genetic studies (Al-Zahery, et al., 2011; Al-Zahery, et al., 2003; Farzad, et al., 2013; Ozbas-Gerceker, Bozman, Arslan, & Serin, 2013; Tomas, Diez, Moncada, Borsting, & Morling, 2013), although my findings differ from Banoei, et al., (2008) who found, using autosomal VNTRs, that Iraqi Arabs were more similar to Kuwaiti Arabs than Iranian Arabs.

Analysis of Molecular Variance

Analysis of Molecular Variance (AMOVA) is a method of estimating population differentiation and can be used to test whether the genetic diversity in each subpopulation is significantly different from that of the total population. A randomization test with 999 permutations were used to obtain *p*-values for the test. If the null hypothesis of no variation between subpopulation is true, then the expectation would be for the sigma-value of the dataset to fall within the non-significance level of the distribution of sigma-values for the randomized permutations.

The results of the AMOVA on this dataset showed that the sigma-value for the variation within samples was significantly lower than the randomized distribution while the sigma-values for the variation between samples and between subpopulations was significantly higher than the randomized distribution (Table 4 and Figure 7). This would lead to rejection of the null hypothesis of no population stratification. Based on the AMOVA, it appears that there is genetic differentiation between the six cities despite the amount of variation that is between the cities being very low (0.15%).

Discriminant Analysis of Principal Components

Discriminant analysis (DA) is a statistical analysis that classifies data into *a priori* groupings based on the measured characteristics of data assigned to those groups. Discriminant

analysis of principal components (DAPC) performs DA after the dataset has been transformed into PCs in order to meet the assumptions of DA. Two DAPCs were performed on this dataset: one in which individuals were assigned to genetic clusters formed by *k*-means clustering and the other in which the city the sample was taken from was the assigned group.

Genetic clustering was done on $k = 1$ through $k = 40$ groupings with the $k = n$ that had the lowest Bayesian Information Criterion (BIC) score chosen for use in the DAPC. For this dataset, that was $k = 8$ (Table 5 and Figure 9). Based on this, it appears that not only is there genetic substructure within the population but there are more substructures than just six, the number of cities sampled. The plot of the DAPC showed overlap of the eight genetic clusters but with clear separation of centers (Figure 10). A χ^2 contingency test was performed to test whether an individual's city of residence was independent of the genetic cluster they were assigned to. The results indicate that the genetic clusters are not independent of city. While all genetic clusters are represented in all cities, they are not randomly distributed and it does appear that some of the genetic substructure is organized according to city/region.

DAPC performed with the cities forming the *a priori* groups was able to correctly assign individuals to their city 48.73% of the time, with the lowest correct assignment rate belonging to Wasit (32.50%) and the highest belonging to Baghdad (69.21%) (Table 9 and Figure 14). However, even 32.50% is greater than would be expected to occur randomly indicating that the cities are genetically distinct enough from each other for better than random assignment to the correct city.

Baghdad has the most correct assignments despite being the city with the lowest Nei's genetic distances. This occurred because the model had difficulty distinguishing the other cities from Baghdad and over-assigned individuals to that city. As the number of PCs retained

decrease, correct assignment to Baghdad increases while correct assignment to the other cities decreases (Figure 15). This behavior supports the conclusion that Baghdad is representative of the country. It is a more ethnically mixed population and is a metropolis where people are likely to either move to or move out off. Additionally, as the capital and largest city, it is possible that people traveled to Baghdad for use of its hospitals which is where samples were collected.

This is reinforced by the DAPC plot which places Baghdad in the exact center where it overlaps with all the other cities' groupings (Figure 13). Basra is next closest to center while Diyala, Najaf, and Anbar are on the surrounding edges. Wasit is on the outside, but closer to the center. This sort of distribution is consistent with what was seen in the PCA plot of the cities. Overall, there is differentiation among the cities, but not as spread out as the genetic clusters formed by k-means clusters. It appears that there is some stratification among the cities/regions, but that is not the main substructure as each genetic cluster is represented in each city and there are more genetic clusters than there are cities sampled.

Multidimensional Scaling

Multidimensional scaling (MDS) is method for visualizing similarity between groups, like PCA, but the two methods differ in in how this similarity is measured. MDS was performed on the allele frequency data for the country of Iraq to compared it to countries around the world. The expectation was that Iraq would cluster with other Middle Eastern countries and be separate from countries from Europe, Asia, and Africa. This was done to ensure that there was no obvious problem with the dataset, such as sequencing error, that would result in the wrong placement. The MDS plot was as expected (Figure 16). Iraq clustered with the other Middle Eastern countries, and closer to Turkey and Iran as was seen in the PCA plot (Figure 6). All the Middle Eastern countries plotted closer to the countries from Europe than from either Asia or Africa,

which is consistent with what has been found in previous studies (Abu-Amero, et al., 2009; Al-Zahery, et al., 2011; Al-Zahery, et al., 2003; Cavalli-Sforza, Menozzi, & Piazza, 1994; Chiaroni, et al., 2010; Nebel, et al., 2001; Richards, et al., 2000). The 3D MDS plot plotted Turkey closest to countries from Europe, followed by Iran, and then Iraq. These three countries formed a group (Figure 17). Of the Middle Eastern countries, Saudi Arabia plotted furthest from European countries, and Syria and Kuwait formed a group that plotted between Saudi Arabia and Turkey, Iran, and Turkey.

Forensic Applications of Population Structure

For the establishment of a forensic DNA database for Iraq, allele frequencies, heterozygosities (He), homozygosities (Ho), polymorphism information content (PIC), matching probabilities (MP), power of discrimination (PD), power of exclusion (PE), and the typical paternity index (TPI) were calculated for each locus for each city. Heterozygosity can range from 0.0 (no heterozygotes in the collected samples) to 1.0 (all the samples are heterozygotes). PIC can range from 0.0 to 1.0 with 0.0 meaning that there is no allelic variation among the samples and 1.0 being total allelic variation. MP can range from 0.0 (zero likelihood of a random match) to 1.0 (complete likelihood of random match). PD ranges from 0.0 to 1.0 with 0.0 being zero power to discriminate between individuals and 1.0 being absolute power to discriminate between individuals. PE ranges from 0.0 (no power to exclude an individual) to 1.0 (complete power to exclude an individual). TPI is a ratio that that starts from 0.0 and continues on up. The value is the number of times more likely it is that the tested man is the true biological father than a randomly selected man from the same population. MP, PD, PE, and TPI will vary from genotype to genotype; what has been calculated here are the average values from the dataset.

For a locus to be useful for forensic and paternity testing, it should have a large He, small Ho, large PIC, large PD, and large PE. The average values across all 15 loci are: He = 0.779, Ho = 0.221, PIC = 0.77, PD = 0.927, and PE = 0.563 (Tables 10 and 11). The two measures listed on forensic or paternity reports, MP and TPI, are multiplied across all tested loci for a combined MP (CMP) and combined PI (CPI). CMP is expressed in forensic reports as its inverse: 1 individual in $\frac{1}{CMP}$ individuals will have a matching genotype. There are roughly 7.5 billion people on this planet. For the CMP to be expressed as 1 in 7.5 billion or more, the average MP across 15 loci would need to be 0.173 or less. The average MP across 15 loci for this dataset was 0.073, and the average CMP was 1.227×10^{-18}. On average, 1 in 8.152×10^{17} individuals will have a matching genotype. Paternity reports will provide a probability of paternity which is calculated as $\frac{CPI}{CPI+1}$. To have a probability of paternity greater than or equal to 99.0%, the CPI must be 100 or more. For 15 loci, the average TPI would then need to be 1.43 or greater. The average TPI across the 15 loci was 2.36 with an average CPI of 289,622.

According to these measures, a matching genotype at all 15 loci for a member of the Iraqi population would, on average, have a probability of occurring randomly on an order of 10^{-18} which can be taken as an exact match. For paternity tests, the probability that a member of the Iraqi population matching a single allele at all 15 loci is the true biological father is 99.9997%, which can be taken as an exact match. It can be concluded that these loci are appropriate for forensic and paternity testing for the sampled population.

Limitations of the Study

The individuals that were sampled for this study were patients and laboratory workers at hospitals and private laboratories, so this is not a random sample of the Iraqi population. The data that were collected were de-identified, meaning that no information was available on the individuals' gender, religion, ethnicity, age, or place of birth. Individuals are only identified by the city in which their sample was collected. It is likely that many individuals are residing in the city of their birth. A portion of the samples were collected from patients in hospitals, and it is possible that at least some of the individuals traveled to a hospital in different city from their place of residence. Additionally, Baghdad, being the capital and largest city, is likely a destination for migration. Also, even if the city their sample was collected in is the city of their birth, that does not mean that their family has resided in that city for generations.

The null hypothesis was that cities are not subpopulations, but rather all a single population. The alternative hypothesis for these analyses was that each city is a subpopulation of Iraq and the individuals assigned to that city are members of that subpopulation. The results were then interpreted as laying somewhere on a continuum between these two extreme outcomes (completely different vs identical). Incorrect assignment of city to some individuals would result in a decreased ability to detect population stratification among the cities. This was dealt with by doing two DAPCs: one in which individuals were grouped by city (Figure 13) and one in which individuals were organized into genetic clusters (Figure 10). These plots were compared and a χ^2 test for independence examined whether the cities were completely independent of the estimated genetic clusters. This provided evidence for whether there were subpopulations within Iraq and whether those subpopulations were organized by city/region.

Without demographic information for this dataset, it is impossible to say in what way any measured genetic differentiation is organized. They may be organized by ethnicity or religious affiliation, or there may be microdifferentiation among regions of similar ethno-religious background due to isolation. However, educated choices were made based on known demographic information for the regions in which samples were collected.

Chapter 6: Conclusions

Summary of Conclusions

There are five general conclusions that can be made from the results obtained in this study. First, Iraq is similar to other countries in the Middle East, particularly Iran and Turkey, and is more similar to Europe than to either Asia or Africa at these 15 forensic STRs. This conclusion is supported by the multidimensional scaling (MDS) and principal components analysis (PCA) plots. The 2D MDS plotted Iraq with other Middle Eastern countries, and these were all plotted closer to Europe (Figures 16). This is also supported by the literature, in which previous studies found genetic similarities between Iraq and Europe (Abu-Amero, et al., 2009; Al-Zahery, et al., 2003; Al-Zahery, et al., 2011; Cavalli-Sforza, Menozzi, & Piazza, 1994; Chiaroni, et al., 2010; Nebel, et al., 2001; Richards, et al., 2000).

The PCA plotted Iraq closest to Iran and Turkey and furthest from Syria and Saudi Arabia (Figure 6). The 3D MDS plot further separated out the Middle Eastern countries with Iran, Iraq, and Turkey forming a group with Saudi Arabia furthest from them and Syria and Kuwait forming a group in between them (Figure 17). Previous studies also found Iraq to be genetically similar to Iran and Turkey (Al-Zahery, et al., 2011; Al-Zahery, et al., 2003; Farzad, et al., 2013; Ozbas-Gerceker, Bozman, Arslan, & Serin, 2013; Tomas, Diez, Moncada, Borsting, & Morling, 2013) and that countries in the Levant are genetically dissimilar to those on the Arabian Peninsula (Badro, et al., 2013; El-Sibai, et al., 2009).

The second conclusion that can be drawn is that there is a fair amount of genetic diversity within Iraq. The best fit when inferring the number of genetic clusters in the population was eight genetic clusters (Table 5 and Figure 9). This is supported by previous studies that have found the influence of Asia, Africa, and Europe in the Arab population of Iraq (Al-Zahery, et al.,

2003; Al-Zahery, et al., 2011; Richards, et al., 2003). In addition to admixture, there are also several distinct ethnicities living within Iraq: Arabic, Kurdish, Turkmen, Assyrian, Jewish populations are all found in Iraq as well as other minority populations. It would be expected for there to be genetic diversity in a country made up of so many populations.

Thirdly, genetic structure does differ between the six cities. The analysis of molecular variance (AMOVA) revealed significantly more genetic variation between the cities than would be expected to occur randomly (Table 4 and Figure 7). Both the PCA and the discriminant analysis of principal components (DAPC) showed dispersal among the cities (Figures 5 and 13). The DAPC also assigned individuals to the correct city more often than would happen randomly (Table 9 and Figure 14). Finally, a χ^2 test of independence found that the genetics clusters were not independent of city.

Baghdad plotted closest to the middle of the PCA and DAPC plots and had the lowest Nei's genetic distances from the other cities (Table 3). It is the largest city and capital of Iraq, and the largest sample size was taken from Baghdad. The most diversity of ethnicity is found in that city. It is a good representation of the rest of the country. Anbar, however, plotted furthest from the other cities and had the highest genetic distances from the other cities. The reason that Anbar is genetically distinct from the other cities may be because Anbar is in a region of Iraq where the population is mainly Sunni while the other cities are in a region where the population is mainly Shia (Figure 18). Baghdad is the city that Anbar has the smallest genetic difference with, and Baghdad's proportions of Sunni and Shia is more mixed than the other cities. It is possible that there is microdifferentiation underway between Sunni and Shia Arabs due marriage being confined within these religious groups.

The fourth conclusion drawn from this data is that, while genetic structure does differ between the cities, most of the genetic differentiation is not between the cities. When comparing the two DAPC plots, the genetic clusters are more dispersed on their plot (Figure 10) than the cities are in their plot (Figure 13). Each genetic cluster was represented in all cities (Table 6) so, although genetic cluster was not independent of city, they are not confined to cities indicating that there has been movement. The overall population, with all cities combined, was found to be in Hardy-Weinberg equilibrium at these loci (Table 1) and the F_{ST} was small (Table 2). This seems to be because city is not the primary organization for the genetic clusters.

The primary structure for the genetic differentiation is according to genetic cluster and these were represented in all cities. There is a secondary structure according to city which may be due to first-cousin marriage being quite common (approximately one-quarter) in Iraq (Al-Allawi, et al., 2015). Individuals are, mainly, marrying within their own ethno-religious population and then, to some degree, within their family who are likely to be living in the same region. This may result in differentiation between regions.

The fifth conclusion drawn from these analyses is that these 15 STRs are appropriate for use in forensics and paternity testing. They averaged a large heterozygosity, polymorphism information content, power of discrimination, and power of exclusion as well as a low homozygosity and matching probability which make them useful for these purposes. At these loci, a matching genotype will occur, on average, in 1 in 8.152×10^{17} individuals. For paternity tests, the average paternity probability for a matching profile will be 99.9997%. For both measures, this can be taken as an exact match.

Future Studies

This study has raised questions that can be explored in future research. A major question is whether the genetic clusters inferred represent known categories. Are the genetic clusters different ethnicities? Do they reveal microdifferentiation within an ethnicity along religious lines with a split occurring between Sunni and Shia Arabs? Are there varying amounts of admixture within the Iraqi population? Or some combination of these? Another important question is what is causing the secondary differentiation between the cities. Could it be due to marrying within families?

These are all testable hypotheses. To clarify the causes behind the genetic differentiation, any future studies will need to collect pertinent information from participants. This includes ethnicity, religious affiliation, place of birth, parents' places of birth, grandparents' place of birth, and incidence of consanguineous marriage within the family. Considering these questions will also improve the forensic database that was established in this study. Genetic structure is what make forensic analysis possible. Right now, the database is set up with allele frequencies for each city, which do differ in genetic structure. However, since the primary differentiation is between genetic clusters, knowing what those represent and obtaining allele frequencies for those groups will be beneficial for any forensic analyses done with this database.

Works Cited

Abdin, L., Shimada, I., Brinkmann, B., & Hohoff, C. (2003). Analysis of 15 short tandem repeats reveals significant differences between the Arabian populations from Morocco and Syria. *Legal Medicine, 5*, S150-S155.

Abu-Amero, K., Hellani, A., González, A. M., Larruga, J. M., Cabrera, V. M., & Underhill, P. A. (2009). Saudi Arabian Y-chromosome diversity and its relationship with nearby regions. *BMC Genetics, 10*, 59.

Al-Allawi, N. A., Nasir, A. S., Al-Doski, A. A., Markous, R. S., Amin, K. A., Eissa, A. A., . . . Hamamy, H. (2015). Premarital screening for hemoglobinopathies: Experience of a single center in Kurdistan, Iraq. *Public Health Genomics, 18*(2), 97-103.

Alenizi, M., Goodwin, W., Ismael, S., & Hadi, S. (2008). STR data for the AmpFℓSTR® Identifiler® loci in Kuwaiti population. *Legal Medicine, 10*(6), 321-325.

Alves, C., Gusmão, L., López-Parra, A. M., Mesa, M. S., Amorim, A., & Arroro-Pardo, E. (2005). STR allelic frequencies for an African population sample (Equatorial Guinea) using AmpFISTR Identifiler and Powerplex 16 kits. *Forensic Science International, 148*(2-3), 239-242.

Al-Zahery, N., Pala, M., Battaglia, V., Grugni, V., Hamod, M. A., Kashani, B. H., . . . Semino, O. (2011). In search of the genetic footprints of Sumerians: A survey of Y-chromosome and mtDNA variation in the Marsh Arabs of Iraq. *BMC Evolutionary Biology, 11*, 288.

Al-Zahery, N., Semino, O., Benuzzi, G., Passarino, G., Torroni, A., & Santachiara-Benerecetti, A. (2003). Y-chromosome and mtDNA polymorphisms in Iraq: A crossroad of the early human dispersal and of post-Neolithic migrations. . *Molecular Phylogenetic Evolutions, 28*, 458-472.

Associated Press. (2007, May 10). *Iraq Bill Demands U.S. Troop Withdraw*. Retrieved from Fox News: http://www.foxnews.com/story/2007/05/10/iraq-bill-demands-us-troop-withdraw.html

Badro, D. A., Douaigy, B., Haber, M., Youahnna, S. C., Salloum, A., Ghassibe-Sabbagh, M., . . . The Genographic Consortium. (2013). Y-Chromosome and mtDNA Genetics Reveal Significant Contrasts in Affinities of Modern Middle Eastern Populations with European and African Populations. *PloS one, 9*(1), e54616.

Balloux, F., & Lugon-Moulin, N. (2002). The estimation of population differentiation with microsatellite markers. *Molecular Ecology, 11*(2), 155-165.

Banoei, M. M., Chaleshtori, M. H., Sanati, M. H., Shariati, P., Houshmand, M., Majidizadeh, T., . . . Golalipour, M. (2008). Variation of DAT1 VNTR Alleles and genotypes among old ethnic groups in Mesopotamia to the oxus region. *Human Biology, 80*(1), 73-81.

BBC News. (2006, July 13). Iraq province power transferred. Retrieved from BBC News: http://news.bbc.co.uk/2/hi/middle_east/5175478.stm

BBC News. (2006, August 28). *Fierce battles in south Iraq city.* Retrieved from BBC News: http://news.bbc.co.uk/2/hi/middle_east/5293278.stm

BBC News. (2006, December 30). *Saddam Hussein executed in Iraq*. Retrieved from BBC News: http://news.bbc.co.uk/2/hi/middle_east/6218485.stm

BBC News. (2007, February 21). *Blair announces Iraq troops cut.* Retrieved from BBC News: http://news.bbc.co.uk/2/hi/uk_news/politics/6380933.stm

BBC News. (2009, June 30). *US Soldiers leave Iraq's cities.* Retrieved from BBC News: http://news.bbc.co.uk/2/hi/middle_east/8125547.stm

BBC News. (2012, December 18). *Iraqi President Jalal Talabani 'in coma after stroke'.* Retrieved from BBC News: http://www.bbc.com/news/world-middle-east-20766435

BBC News. (2012, December 28). *Iraq Sunni protests in Anbar against Nouri al-Maliki*. Retrieved from BBC News: http://www.bbc.com/news/world-middle-east-20860647

BBC News. (2013, January 2). *Protests engulf west Iraq as Anbar rises against Maliki*. Retrieved from BBC News: http://www.bbc.com/news/world-middle-east-20887739

BBC News. (2014, June 18). *Iraq crisis: Battle grips vital Baiji oil refinery.* Retrieved from BBC News: http://www.bbc.com/news/world-middle-east-27897648

BBC News. (2014, September 9). *Haider al-Abadi: A new era for Iraq?* Retrieved from BBC News: http://www.bbc.com/news/world-middle-east-28748366

Beleza, S., Alves, C., Reis, F., Amorim, A., Carracedo, A., & Gusmão, L. (2005). 17 STR data (AmpF/STR Identifiler and Powerplex 16 System) from Cabinda (Angola). *Forensic Science International, 141*(2-3), 193-196.

Borg, L., & Groenen, P. (1997). *Modern Multidimensional Scaling, Theory and Applications.* . New York, NY: Springer-Verlag.

Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics, 32*, 314-331.

Brenner, C., & Morris, J. (1990). Paternity index calculations in single locus hypervariable DNA probes: Validation and other studies. In *Proceedings for the International Symposium on Human Identification* (pp. 21-53). Madison, WI: Promega Corporation.

Castle, W. E. (1903). The laws of Galton and Mendel and some laws governing race improvement by selection. *Proceedings of the American Academy of Arts and Sciences, 35*, 233-242.

Cavalli-Sforza, L., Menozzi, P., & Piazza, A. (1994). *The history and geography of human genes.* Princeton, NJ: Princeton University Press.

Chessel, D., Dufour, A. B., & Thioulouse, J. (2004). The ade4 package-I-One-table methods. *R News, 4*(1), 5-10.

Chiaroni, J., King, R. J., Myre, N. M., Henn, B. M., Ducourneau, A., Mitchell, M. J., . . . Underhill, P. A. (2010). The emergence of Y-chromosome haplogroup J1e among Arabic-speaking populations. *European Journal of Human Genetics, 18*, 348-353.

Cline, L. E. (2000). The Prospects of the Shia Insurgency Movement in Iraq. *The Journal of Conflict Studies, 20*(2).

CNN. (2006, November 12). *Iraqi leader pushes Cabinet changes as bombs kill dozens.* Retrieved from CNN: http://www.cnn.com/2006/WORLD/meast/11/12/iraq.main/index.html

CNN. (2011, December 18). *Deadly Iraq war ends with exit of last US troops*. Retrieved from CNN: http://edition.cnn.com/2011/12/17/world/meast/iraq-troops-leave/index.html

CNN. (2014, August 11). *Iraq's Nuri al-Maliki digs in as President nominates new Primer Minister.* Retrieved from CNN: http://www.cnn.com/2014/08/11/world/meast/iraq-crisis/index.html?hpt=hp_t2

Comas, D., Calafell, F., Bendukidze, N., Fañanás, L., & Bertranpetit, J. (2000). Georgian and Kurd mtDNA sequence shows a lack of correlation between languages and female genetic lienages. *American Journal of Physical Anthropology, 112*, 5-16.

Decorte, R., Gilissen, A., & Cassiman, J. J. (2003). Allele frequency data for 15 STR loci (AMPFISTR®SGM plusTM and AmpFISTR® ProfilerTM) in the Belgian population. *International Congress Series, 1239*, 219-222.

Dharmarajan, G., Beatty, W. S., & Rhodes Jr, O. E. (2012). Heterozygote deficiencies caused by a Wahlund effect: Dispelling unfounded expectations. *The Journal of Wildlife Management, 77*(2), 226-234.

Donaldson, M., Tucket, S., & Grant, D. (1980). Recessively inherited growth-hormone deficiency in a family from Iraq. *Journal of Medical Genetics, 17*(4), 288-290.

Dray, S., & Dufour, A. B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software, 22*(4), 1-20.

Dray, S., Dufour, A. B., & Chessel, D. (2007). The ade4 package-II: Two-table and K-table methods. *R News, 7*(2), 47-52.

El-Sibai, M., Platt, D. E., Haber, M., Zue, Y., Youhanna, S. C., Wells, R. S., . . . The Genographic Consortium. (2009). Geographical Structure of the Y-chromosomal Genetic Landscape of the Levant: A coastal-inland contrast. *Annals of Human Genetics, 73*(6), 568-581.

Elad-Altman, I. (2007). The Sunni-Shi'a Conversion Controversy. *Current Trends in Islamist Ideology, 5*, 1-10.

Engels, W. R. (2009). Exact tests for Hardy-Weinberg proportions. *Genetics, 183*(4), 1431-1441.

Eriksen, T. H. (2010). *Ethnicity and Nationalism: Anthropological Perspectives*. New York City, NY: Palgrave MacMillan.

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data. *Genetics, 131*, 479-491.

Farzad, M. S., Tomas, C., Borsting, C., Zeinali, Z., Malekdoost, M., Zeinali, S., & Morling, N. (2013). Analysis of 49 autosomal SNPs in three ethnic groups from Iran: Persians, Lurs and Kurds. *Forensic Science International-Genetics, 7*(4), 471-473.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Human Genetics, 7*(2), 179-188.

Fisher, R. A., & Yates, F. (1943). *Statistical Tables: For Biological, Agricultural and Medical Research.* London, UK: Oliver & Boyd.

Foster, B. R., & Foster, K. P. (2009). *Civilizations of Ancient Iraq.* Princeton, NJ: Princeton University Press.

Goldschmidt, E., Fried, K., Steinberg, A. G., & Cohen, T. (1976). Karaite community of Iraq in Israel - genetic study. *American Journal of Human Genetics, 28*(3), 243-252.

Google Maps. (2017, April 14). *Iraq*. Retrieved from https://www.google.com/maps/place/Iraq

Guardian, The. (2014, August 11). *Kerry slaps down Maliki after he accuses Iraqi president of violating constitution.* Retrieved from The Guardian: https://www.theguardian.com/world/2014/aug/11/us-iraqi-maliki-accuses-president

Guo, S. W., & Thompson, E. A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics, 48*, 361-372.

Haldane, J. B. (1954). An exact test for randomness of mating. Journal of Genetics. *Journal of Genetics, 52*(3), 631-635.

Hamamy, H., & Al-Hakkak, Z. (1989). Consanguinity and reproductive health in Iraq. *Human Heredity, 29*(5), 271-275.

Hamamy, H., Makrythanasis, P., Al-Allawi, N., Muhsin, A., & Antonarakis, S. (2014). Recessive thrombocytopenia likely due to a homozygous pathogenic variant in the FYB gene: Case report. *BMC Medical Genetics, 1*, 135.

Hardy, H. G. (1908). Mendelian proportions in a mixed population. . *Science, 28*, 49-50.

Hartigan, J. A., & Wong, M. A. (1979). A K-Means Clustering Algorithm. *Applied Statistics, 28*, 100-108.

Hashiyada, M. (2000). Short tandem repeat analysis in Japanese population. *Electrophoresis, 21*(2), 347-350.

Henningsen, E., Svendsen, M., Lildballe, D., & Jensen, P. (2014). A novel mutation in the EDAR gene causes severe autosomal recessive hypohidrotic ectodermal dysplasia. *American Journal of Medical Genetics Part A, 164*(8), 2059-2061.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics, 6*(2), 65-70.

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structure populations: defining, estimating and interpreting FST. *Nature Reviews Genetics, 10*(9), 639-650.

Hotelling, H. (1933a). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*(6), 417-441.

Hotelling, H. (1933b). Analysis of a complex of statistical variables into principal components (Continued from September issue). *Journal of Educational Psychology, 24*(7), 498-520.

Hourani, A. (1991). *A History of the Arab Peoples*. Cambridge, MA: Belknap Press of Harvard University Press.

Jombart, T. (2008). adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics, 24*(11), 1403.

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics, 27*(21), 3070-3071.

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genetics, 11*, 94.

Kamvar, Z. N., Brooks, J. C., & Grünwald, N. J. (2015). Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in Genetics, 6*, e208.

Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr: An R package for genetic analysis of populations with clonal, partial clonal, and/or sexual reproduction. *PeerJ, 2*, e281.

Kirmanj, S. (2013). *Identity and Nation in Iraq.* Boulder, CO: Lynne Rienner Publishers.

Lachenbruch, P. A., & Goldstein, M. (1979). Discriminant Analysis. *Biometrics, 35*(1), 69-85.

Levene, H. (1949). On a matching problem arising in genetics. *The Annals of Mathematical Statistics, 20*(1), 91-94.

Marsaglia, H. A. (2003). Random number generators. *Journal of Modern Applied Statistical Methods, 2*, 2-13.

Mehta, C. R., & Patel, N. R. (1983). A Network Algorithm for Performing Fisher's Exact Test in r x c Contingency Tables. *Journal of the American Statistical Association, 78*(382), 427-434.

Nebel, A., Filon, D., Brinkmann, B., Majumder, P., Faerman, M., & Oppenheim , A. (2001). The Y chromosome pool of Jews as a part of the genetic landscape of the Middle East. . *American Journal of Human Genetics, 69*(5), 1095-1112.

Nei, M. (1972). Genetic distances between populations. *American Naturalist, 106*, 283-292.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics, 23*, 341-369.

Nei, M., & Takezaki, N. (1996). The Root of the Phylogenetic Tree of Human Populations. *Molecular Biology and Evolution, 13*(1), 170-177.

NY Times. (2006, October 20). *Attack on Iraqi City Shows Militia's Power*. Retrieved from NY Times: http://www.nytimes.com/2006/10/20/world/middleeast/21iraqcnd.html?ex=1318996800&en=a542d37a1dff56f9&ei=5088&partner=rssnyt&emc=rss

NY Times. (2017, September 2). *ISIS Is on Its Heels, but Fighting to the Death*. Retrieved from NY Times: https://www.nytimes.com/2017/09/02/world/middleeast/isis-iraq-fight.html?rref=collection%2Ftimestopic%2FIraq&action=click&contentCollection=world&region=stream&module=stream_unit&version=search&contentPlacement=1&pgtype=collection

NY Times. (2017, September 26). *Iraq Orders Kurdistan to Surrender Its Airports.* Retrieved from NY Times: https://www.nytimes.com/2017/09/26/world/middleeast/iraq-kurds-independence.html

Osman, A. E., Alsafar, H., Tay, G. K., Theyab, J., Mubasher, M., Eltayeb-El Sheikh, N., . . . El Ghazali, G. (2015). Autosomal Short Tandem Repeat (STR) Variation Based on 15 Loci

in a Population from the Central Region (Riyadh Province) of Saudi Arabia. *Journal of Forensic Research, 6*, 267.

Ozbas-Gerceker, F., Bozman, N., Arslan, A., & Serin, A. (2013). Population Data for 17 Y-STRs in Samples from Southeastern Anatolia Region of Turkey. *International Journal of Human Genetics, 13*(2), 105-111.

Paradis, E. (2010). pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics, 26*(3), 419-420.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series, 50*(302), 157-175.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, 5*, 559-572.

R Core Team. (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. Retrieved from R Foundation for Statistical Computing: https://www.R-project.org/

Rao, C. R. (1948). The Utilization of Multiple Measurements in Problems of Biological Classification. *Journal of the Royal Statistical Society, Series B (Methodological), 10*(2), 159-203.

Raymond, M., & Rousset, F. (1995). GENEPOP (version 1.2): Population genetics software for exact tests and ecumenicism. *Journal of Heredity, 86*, 248-249.

Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., . . . Bandelt, H. J. (2000). Tracing European founder lineages in the Near Eastern mtDNA pool. *American Journal of Human Genetics, 67*(5), 1251-1276.

Richards, M., Rengo, C., Cruciani, F., Gratrix, F., Wilson, J. F., Scozzari, R., . . . Torroni, A. (2003). Extensive female-mediated gene flow from sub-Saharan Africa into Near Eastern Arab populations. *American Journal of Human Genetics, 72*(4), 1058-1064.

Rousset, F. (2008). Genepop'007: A complete reimplementation of the Genepop software for Windows and Linux. *Molecular Ecology Resources, 8*, 103-106.

Rousset, F., & Raymond, M. (1995). Testing heterozygote excess and deficiency. *Genetics, 140*, 1413-1419.

Roux, G. (1992). *Ancient Iraq* (3rd Edition ed.). London, UK: Penguin Group.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6*(2), 461-464.

Shepard, E. M., & Herrera, R. J. (2006). Iranian STR variation at the fringes of biogeographical demarcation. *Forensic Science International, 158*(2-3), 140-148.

Soetaert, K. (2017). plot3D: Plotting Multi-Dimensional Data. *R Package Version 1.1.1*

Spuler, B. (1994). *A history of the Muslim world.* Princeton, NJ: M. Wiener Publishers.

Stansfield, G. (2007). *Iraq: People, History, Politics.* Malden, MA: Polity Press.

Stansfield, G. (2016). *Iraq: People, History, Politics* (2nd ed.). Malden, MA: Polity Press.

Szczerkowska, Z., Kapińska, E., Wycocka, J., & Cybulska, L. (2004). Northern Polish population data and forensic usefulness of 15 autosomal STR loci. *Forensic Science International, 144*(1), 69-71.

Tereba, A. (1999). Tools for analysis of population statistics. *Profiles in DNA, 2(3)*, 14-16. Promega Corporation.

Time. (2008, January 31). *A year after Bush sent 30,000 additional troops to Iraq, violence is down and al-Qaeda is in retreat. But the gains are still too fragile.* Retrieved from Time: http://content.time.com/time/magazine/article/0,9171,1708843,00.html

Tomas, C., Diez, I. E., Moncada, E., Borsting, C., & Morling, N. (2013). Analysis of 49 autosomal SNPs in an Iraqi population. *Forensic Science International-Genetics, 7*(1), 198-199.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika, 27*(4), 401-419.

Tripp, C. (2002). *A History of Iraq* (2nd ed.). Cambridge, UK: Cambridge University Press.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S.* New York, NY, USA: Spring.

Wahlund, S. (1928). . Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas, 11*, 65-106.

Walker, A. (2015). openxlsx: Read, Write and Edit XLSX Files. *R package version 3.0.0*.

Wang, Z. Y., Yu, R. J., Wang, F., Li, X. S., & Jin, T. B. (2005). Genetic polymorphisms of 15 STR loci in Han population from Shaanxi (NW China). *Forensic Science International, 147*(1), 89-91.

Washington Post. (2006, August 25). *British Leave Iraqi Base; Militia Supporters Jubilant.* Retrieved from Washington Post: http://www.washingtonpost.com/wp-dyn/content/article/2006/08/24/AR2006082401917.html?nav=rss_nation/special

Washington Post. (2006, August 26). *Looters Ransack Base After British Depart.* Retrieved from Washington Post: http://www.washingtonpost.com/wp-dyn/content/article/2006/08/25/AR2006082501315.html?nav=rss_nation/special

Weinberg, W. (1908). On the demonstration of heredity in man. In *Papers on human genetics.*, 1963; 4-15. (S. H. Boyer, Ed.) Englewood Cliffs, NJ: Prentice-Hall.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution, 38*, 1358-1370.

Weir, B. S., & Hill, W. G. (2002). Estimating F-statistics. *Annual Review of Genetics, 36*, 721-750.

World Bulletin. (2014, July 24). *Iraq selects senior Kurdish politician Fuad Masum president.* Retrieved from World Bulletin: http://www.worldbulletin.net/headlines/141288/iraq-selects-fuad-masum-as-new-president

World Factbook, The. (2017, October 13). *Iraq*. Retrieved from Central Intelligence Agency: https://www.cia.gov/library/publications/the-world-factbook/geos/iz.html

Wright, S. (1943). Isolation by Distance. *Genetics, 28*(2), 114-138.

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics, 15*, 323-354.

Yavuz, I., & Sarikaya, A. T. (2005). Turkish Population Data for 15 STR Loci by Multiplex PCR. *Journal of Forensic Sciences, 50*(3), 737-738.